

**UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS  
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO  
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA APLICADA  
NÍVEL DOUTORADO**

**ANA LUIZA TREICHEL VIANNA**

**SEMÂNTICA DE *FRAMES*, HARMONIZAÇÃO TERMINOLÓGICA E  
COMPUTAÇÃO: O uso de *frames* semânticos como princípio organizador para a  
harmonização de termos e conceitos e a representação do conhecimento em *Large  
Language Models***

**São Leopoldo**

**2024**

ANA LUIZA TREICHEL VIANNA

**SEMÂNTICA DE *FRAMES*, HARMONIZAÇÃO TERMINOLÓGICA E  
COMPUTAÇÃO: O uso de *frames* semânticos como princípio organizador para a  
harmonização de termos e conceitos e a representação do conhecimento em *Large  
Language Models***

Tese de Doutorado apresentada como requisito parcial para obtenção do título de Doutora em Linguística Aplicada, pelo Programa de Pós-Graduação em Linguística Aplicada da Universidade do Vale do Rio dos Sinos (UNISINOS).

Orientador(a): Prof. Dr. Sandro José Rigo

Coorientador (a): Profa. Dra. Cátia de Azevedo Fronza

São Leopoldo

2024

V617s

Vianna, Ana Luiza Treichel.

Semântica de frames, harmonização terminológica e computação : o uso de frames semânticos como princípio organizador para a harmonização de termos e conceitos e a representação do conhecimento em large language models / por Ana Luiza Treichel Vianna. – 2024.

321 f. : il. ; 30 cm.

Tese (doutorado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Linguística Aplicada, São Leopoldo, RS, 2024.

Orientador: Dr. Sandro José Rigo.

Coorientadora: Dra. Cátia de Azevedo Fronza.

1. Harmonização de termos e conceitos. 2. Semântica de frames. 3. Frame-based terminology. 4. Large language model (LLM). 5. Knowledge graph. I. Título.

CDU: 801.54:004.8

Catálogo na Publicação (CIP):  
Bibliotecário Alessandro Dietrich - CRB 10/2338

**ANA LUIZA TREICHEL VIANNA**

**“SEMÂNTICA DE FRAMES, HARMONIZAÇÃO TERMINOLÓGICA  
E COMPUTAÇÃO: O USO DE FRAMES SEMÂNTICOS COMO  
PRINCÍPIO ORGANIZADOR PARA A HARMONIZAÇÃO DE  
TERMOS E CONCEITOS E A REPRESENTAÇÃO DO  
CONHECIMENTO EM LARGE LANGUAGE MODELS”**

Tese apresentada como requisito parcial  
para obtenção do título de Doutor, pelo  
Programa de Pós-Graduação em  
Linguística Aplicada da Universidade do  
Vale do Rio dos Sinos - UNISINOS.

**APROVADA EM 3 DE OUTUBRO DE 2024**


**BANCA EXAMINADORA**

**PROFA. DRA. MARIA JOSÉ FINATTO - UFRGS**

**PROF. DR. DIEGO SPADER DE SOUZA - UNISC**

**PROF. DR. RAFAEL KUNST – UNISINOS**

**PROFA. DRA. CÁTIA DE AZEVEDO FRONZA – UNISINOS  
(COORDINADORA)**

Documento assinado digitalmente  
 SANDRO JOSE RIGO  
Data: 03/10/2024 22:12:07-0300  
Verifique em <https://validar.iti.gov.br>

---

**PROF. DR. SANDRO JOSÉ RIGO - UNISINOS  
(ORIENTADOR)**

## **AGRADECIMENTOS**

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

## AGRADECIMENTOS

Essa jornada não teria sido a mesma sozinha. A minha sorte é que muitas pessoas acreditam no meu entusiasmo e nas minhas loucuras acadêmicas e profissionais. São esses devaneios que me trouxeram até aqui. Por isso, é hora de agradecer.

Ao meu esposo, Mirrair, meu maior incentivador. Obrigada por ter aceitado esse desafio comigo e por compreender as minhas ausências, irritações e ansiedades. Obrigada pelos abraços que eu precisava quando o medo e o cansaço batiam. Essa conquista não é só minha, é nossa.

Ao meu orientador, Prof. Dr. Sandro Rigo, por acreditar no meu trabalho e sempre me incentivar a seguir. Muito obrigada por todo o apoio, pelas conversas e pelas longas horas de orientação e de aulas sobre Computação, Inteligência Artificial, Ontologias e todos os tópicos computacionais. Sem o teu encorajamento, eu não teria descoberto o quanto sou fascinada pela Linguística Computacional e como me vejo mais na tecnologia do que na linguagem.

À minha coorientadora, Profa. Dra. Cátia Fronza, que me manteve com os pés na Linguística. Muito obrigada pelas conversas, pelas orientações e por todo o carinho de sempre. Agradeço-te muito por me acolher e por participar desse trabalho.

À Profa. Dra. Rove Chishman. Muito obrigada por todas as nossas conversas, Rove. Obrigada por me aceitar como bolsista de Iniciação Científica e me mostrar o que é pesquisar em meados de 2013. Obrigada por acreditar em mim e me convidar para entrar no Doutorado, em 2020. Todas as tuas orientações foram (e são) essenciais para eu chegar até aqui. Não canso de te agradecer por me mostrar a pesquisa e a Linguística. Depois disso, nunca mais quis deixar de pesquisar.

À minha família, meus pais, Zeno e Rosane, e meu irmão, Mariana e Marcelo. Apesar de não entenderem a minha pesquisa (“Eu não consigo explicar para o pessoal o

que tu estuda”, GP... o quê?”, “Harmonizar? Mas para quê fazer isso?”), nem o meu trabalho como pesquisadora e linguista (“Mas tu só estuda, certo?”, “Isso é trabalho? Tu fica o tempo todo no computador...”, “Ah, tu é estagiária...”), sempre me apoiaram e me incentivaram a seguir em frente. Obrigada por estarem sempre ao meu lado. Amo vocês!

Aos colegas do grupo de pesquisa VLHSem, Eduardo e Mikaela, por me escutarem falar da tese (incansavelmente) em muitas reuniões. Obrigada por me auxiliarem nesse período, por compartilharem ideias e soluções e contribuírem para que esta tese tomasse forma. Agradeço, também, ao Thomas, ao Douglas, à Maitê, à Carolina, ao Ariel e ao Felipe, parte integrante do VLHSem, por me desafiarem neste trabalho, questionando-me sobre as aplicações da harmonização e me fazendo refletir constantemente sobre o tema.

À Izabela, minha confidente e parceira de Pós-Graduação. Obrigada por me ouvir com toda a atenção e com todo o amor que só tu tens.

Ao meu fruto, minha eterna estrelinha.

À CAPES pela bolsa concedida para a realização desta pesquisa.

“It was this combination of passion and perseverance that made high achievers special. In a word, they had grit.”

(Angela Duckworth)

“However, artificial intelligence can and should be studied in ways that are not strictly technical. It is important for us to realize how this science is related to the hopes (and fears) of humanity. To do this we must try to understand people, not just machines. If artificial intelligence is to be developed beneficially, it will have to become one of our most humanistic sciences.”



## RESUMO

Esta tese tem por objetivo investigar o papel da Semântica de *Frames* como princípio organizador de uma base de dados *Enterprise Resource Planning* (ERP), contribuindo para a harmonização terminológica, bem como automatizar tal tarefa em um *Large Language Model* (LLM). Com a transformação digital e o uso de *Large Language Models* (LLMs), como o ChatGPT (OpenAI, 2024), muitas atividades que, anteriormente, eram realizadas manualmente estão sendo automatizadas. No contexto empresarial, para se gerir uma empresa, todo o processamento de dados era feito separadamente. Hoje em dia, contamos com o *software Enterprise Resource Planning* (ERP), que integra todas as áreas de uma instituição a fim de permitir que se compartilhem os dados de todos os setores corporativos, desde Recursos Humanos até Logística. A fim de tornar a comunicação mais eficiente e evitar inconsistências terminológicas, a *International Organization for Standardization* (ISO) desenvolveu a recomendação ISO 860:2007, referente à harmonização de termos e conceitos, que busca harmonizar a terminologia de um domínio. Para essa tarefa, a ISO se baseou nos pressupostos de Wüster (1968, 1974) e da Teoria Geral da Terminologia (Felber, 1979), não considerando os aspectos contextuais, cognitivos e variacionais em sua composição, nem possibilitando que se harmonizem termos de dois ou mais domínios concomitantemente. Em se tratando da terminologia de uma empresa, muitas vezes, utilizamos diferentes termos para nos referirmos a um mesmo conceito em áreas distintas. Para solucionar esse impasse terminológico, em nível teórico, buscaram-se subsídios na Semântica de *Frames* (Fillmore, 1967, 1976, 1982, 1985) para lidar com o mapeamento do ERP e dos setores que o compõem. Para o estudo terminológico, utilizou-se a *Frame-based Terminology* (Faber, 2012, 2014, 2015) para analisar e organizar a terminologia do *software*. Com essa fundamentação, foi proposta uma abordagem semântico-terminológica para a harmonização de termos e conceitos. Em nível prático, arquitetou-se um *Knowledge Graph* com *frames* semânticos do ERP e se automatizou a metodologia de harmonização em um LLM, elaborando diferentes estratégias de *prompt* para que o modelo fizesse uma harmonização considerando a proposta desenvolvida na tese. Como metodologia, utilizou-se a Linguística de *Corpus* para a compilação dos *corpora* de pesquisa a fim de extrair a terminologia e selecionar os termos a serem harmonizados. Após a seleção de termos, analisaram-se três conjuntos terminológicos e se propôs uma harmonização manual para cada grupo.

Posteriormente, utilizaram-se dois modelos de linguagem, ChatGPT-4 e ChatGPT-4 32k, para a automatização da proposta semântico-terminológica para harmonização de termos e conceitos. Com os experimentos realizados neste trabalho, observou-se que a Semântica de *Frames* serve como princípio organizador para o ERP e contribui para a harmonização, fornecendo mais dados linguísticos para a seleção do termo principal e suas variantes. No que se refere à harmonização automática em um LLM, percebeu-se que o modelo em temperatura 1 apresentou uma harmonização similar a realizada manualmente, uma vez que considerou todas as informações contextuais provenientes de definições terminológicas e da representação do conhecimento combinadas com a sua criatividade. Observou-se que uma base semântico-terminológica qualifica a tarefa de harmonização, proporcionando mais contexto e conteúdo linguístico. Além disso, evidenciou-se que é possível automatizar a harmonização em um LLM, obtendo resultados similares à harmonização manual. Por fim, salientamos que a combinação entre Semântica de *Frames*, *Frame-based Terminology* e Inteligência Artificial não só possibilitou a elaboração da estrutura conceitual do ERP, como também contribuiu para a organização da terminologia do *software*, oferecendo um tratamento mais robusto à harmonização de termos e conceitos capaz de automatizar tal tarefa em um *Large Language Model* de forma mais eficiente e precisa.

Palavras-chave: harmonização de termos e conceitos; Semântica de *Frames*; *Frame-based Terminology*; *large language model*; *knowledge graph*.

## ABSTRACT

This PhD dissertation aims to investigate the role of Frame Semantics as an organizing principle for an ERP database, contributing to terminological harmonization, as well as automating this task in a Large Language Model. With digital transformation and the use of Large Language Models (LLMs), such as ChatGPT (OpenAI, 2024), many activities that were previously performed manually are now being automated. In the business context, data processing was once carried out separately to manage a company. Today, we have Enterprise Resource Planning (ERP) software, which integrates all areas of an institution to enable data sharing across all corporate sectors, from Human Resources to Logistics. To make communication more efficient and avoid terminological inconsistencies, the International Organization for Standardization (ISO) developed the ISO 860:2007 recommendation on the harmonization of terms and concepts, which seeks to harmonize terminology within a domain. For this task, ISO relied on the assumptions of Wüster (1968, 1974) and the General Theory of Terminology (Felber, 1979), which do not take contextual, cognitive, and variational aspects into account, nor allow for the harmonization of terms from two or more domains simultaneously. Regarding company terminology, different terms are often used to refer to the same concept in different areas. To address this terminological issue, at a theoretical level, Frame Semantics (Fillmore, 1967, 1976, 1982, 1985) was used to support the mapping of the ERP and its components. For terminological study, Frame-based Terminology (Faber, 2012, 2014, 2015) was utilized to analyze and organize the software's terminology. With this foundation, a semantic-terminological approach for the harmonization of terms and concepts was proposed. On a practical level, a Knowledge Graph was designed with semantic frames of the ERP, and the harmonization methodology was automated in an LLM by developing different prompt strategies for the model to perform harmonization according to the thesis' proposal. Corpus Linguistics was used as the methodology to compile research corpora, allowing for the extraction of terminology and the selection of terms to be harmonized. After selecting the terms, three terminological sets were analyzed, and manual harmonization was proposed for each group. Subsequently, two language models, ChatGPT-4 and ChatGPT-4 32k, were used to automate the semantic-terminological proposal for harmonizing terms and concepts. The experiments carried out in this study showed that

Frame Semantics serves as an organizing principle for ERP and contributes to harmonization, providing more linguistic data for selecting the main term and its variants. Regarding automatic harmonization in an LLM, it was observed that the model at temperature 1 produced harmonization like the one performed manually, as it considered all contextual information from terminological definitions and knowledge representation combined with its creativity. It was observed that a semantic-terminological database enhances the harmonization task by providing more context and linguistic content. Moreover, it was demonstrated that it is possible to automate harmonization in an LLM, achieving results similar to the manual harmonization. Finally, we highlight that the combination between Frame Semantics, Frame-based Terminology and Artificial Intelligence allowed the elaboration of the ERP conceptual structure, as well as it contributed to the organization of the software's terminology, allowing to give a sturdy design to the harmonization of terms and concepts and to automatize such task in a Large Language Model in a more precise and efficient way.

Keywords: harmonization of terms and concepts; Frame Semantics; Frame-based Terminology; large language model; knowledge graph.

## Lista de Figuras

Figura 1 - Frame Transação Comercial	39
Figura 2 - Tela inicial do DiCoEnviro	47
Figura 3 - Frame Evento climático e meteorológico	48
Figura 4 - Frame mudança de temperatura	49
Figura 5 - Tela inicial do EcoLexicon	49
Figura 6 - Termo 'mudança climática'	50
Figura 7 - Tela inicial do Field: dicionário de expressões do futebol	52
Figura 8 - Cenário Gol	52
Figura 9 - Palavra 'Gol de Letra'	53
Figura 10 - Tela inicial e modalidades do Dicionário Olímpico	54
Figura 11 - Modalidade olímpica 'Voleibol'	55
Figura 12 - Cenário Ataque	56
Figura 13 - Palavra 'Bomba'	56
Figura 14 - Tela inicial do Dicionário Paralímpico	57
Figura 15 - Modalidade: Natação paralímpica	58
Figura 16 - Cenário Equipamentos	59
Figura 17 - Palavra 'tapper'	59
Figura 18 - Síntese do capítulo "Linguagem e Cognição: A Semântica de Frames"	62
Figura 19 - Relação entre designação e símbolo de Organização das Nações Unidas	91
Figura 20 - Procedimentos para a harmonização de termos e conceitos	92
Figura 21 - Procedimento de harmonização NBR 13790:1997	93
Figura 22 - Dois sistemas de conceitos diferentes	96
Figura 23 - Exemplo de sistema harmonizado	97
Figura 24 - Síntese do capítulo "As bases terminológicas: teorias da Terminologia e harmonização de termos e conceitos"	105
Figura 25 - Rede neural simples e Rede neural profunda	119
Figura 26 - Subáreas dos estudos da linguagem	122
Figura 27 - Word Embedding em espaço unidimensional	127
Figura 28 - Word Embedding em espaço bidimensional e tridimensional	127
Figura 29 - Área e subáreas da Inteligência Artificial	132
Figura 30 - Linha do tempo da representação do conhecimento	146
Figura 31 - Síntese do capítulo "A interface Linguístico-computacional: Inteligência Artificial, Processamento de Linguagem Natural e Large Language Model"	162
Figura 32 - Dashboard da ferramenta Sketch Engine	173
Figura 33 - Wordlist do corpus gerada pelo Sketch Engine	174
Figura 34 - Lista de keywords gerada pela ferramenta Sketch Engine	174
Figura 35 - Word Sketch gerado pelo Sketch Engine	175
Figura 36 - Recurso Thesaurus da ferramenta Sketch Engine	175
Figura 37 - Ferramenta LancsBox	176
Figura 38 - Recurso GraphColl gerado pela ferramenta LancsBox	177
Figura 39 - Ontologia sobre informação de autor, livros e publicações	179
Figura 40 - Ontologia sobre bebidas	180
Figura 41 - Interligação entre Knowledge Graph	181
Figura 42 - Etapas metodológicas propostas de Müller	182
Figura 43 - Etapas metodológicas para a construção do Knowledge Graph com frames e para a aplicação da harmonização terminológica	187

Figura 44 - Ferramenta CmapTools	189
Figura 45 - Interface do WebProtégé	191
Figura 46 - Interface do Protégé versão desktop	192
Figura 47 - Prompt de cadeia de pensamento	195
Figura 48 - Prompt de árvore de pensamento	196
Figura 49 - Processo de prompt de conhecimento gerado	197
Figura 50 - Modelo de instrução de prompt	199
Figura 51 - Relação entre LLM e Knowledge Graph	200
Figura 52 - Definição dos parâmetros GPT	202
Figura 53 - Prompt GPT-4 32k	207
Figura 54 - Proposta metodológica para a Harmonização de termos e conceitos seguindo a abordagem semântico-terminológica	217
Figura 55 - Síntese do capítulo “Metodologia”	218
Figura 56 - Classe ERP e suas subclasses	222
Figura 57 - Classe ERP, subclasse Human Resource e componentes	223
Figura 58 - Knowledge Graph do setor de Human Resources e alguns recursos e funcionalidades	224
Figura 59 - Componente de Recursos Humanos, recursos e funcionalidades	225
Figura 60 - Recursos HR Tools e Personnel Management e suas funcionalidades	225
Figura 61 - Recurso Personnel Time Management e suas funcionalidades	226
Figura 62 - Instâncias e comentários do recurso de HR Tools	227
Figura 63 - Frames e subframes	228
Figura 64 - Parte do Knowledge Graph de Manufacturing	236
Figura 65 - Subframe HR_Master_Data com anotação do termo 'colaborador'	239
Figura 66 - Termo 'empregado' localizado no Super Frame ERP, instância Cross-Component	240
Figura 67 - Human Resources e subframe Personnel_Management	240
Figura 68 - Anotação Personnel_Management	241
Figura 69 - Subframe Task_and_Resource_Management	242
Figura 70 - Subframe Payroll	242
Figura 71 - Human Resources e Personnel_Management	243
Figura 72 - Anotação Personnel_Management	243
Figura 73 - Anotação instância Business_User	244
Figura 74 - Anotação instância Business_User	244
Figura 75 - Anotação instância HCM_Employee_Central	245
Figura 76 - Anotação instância Business_User	246
Figura 77 - Subframe Treasury_Management	249
Figura 78 - Subframe Central_Purchasing	250
Figura 79 - Subframe Mass_Maintenance_of_Products	250
Figura 80 - Anotação instância Supply_Chain_Cockpit	250
Figura 81 - Subframe Sourcing_and_Contract_Management	251
Figura 82 - Subframe Marketing_Accounting_and_Pricing	251
Figura 83 - Subframe Collections_and_Disbursements	252
Figura 84 - Anotação instância Web_Channel_Sales_Application	252
Figura 85 - Anotação instância Sourcing_OnDemand	253
Figura 86 - Anotação instância Live_Auctions	253
Figura 87 - Anotação instância Dynamic_Pricing_Engine	254
Figura 88 - Subframe Product_Sourcing	256
Figura 89 - Subframe Central_Purchasing	257

Figura 90 - Anotação instância <code>Dynamic_Pricing_Engine</code>	258
Figura 91 - Subframe <code>Material_Requirements_Planning</code>	259
Figura 92 - Anotação instância <code>Manufacturing_Execution</code>	259
Figura 93 - Anotação instância <code>Live_Auctions</code>	260
Figura 94 - Anotação instância <code>Retail</code>	260
Figura 95 - Subframe <code>Sales_and_Distribution</code>	261
Figura 96 - Anotação instância <code>Basic_Integration_with_Extended_Warehouse_Management</code>	261
Figura 97 - Subframe <code>Treasury_Management</code>	262
Figura 98 - Sistematização do capítulo “Análise de dados”	288

## Lista de Quadros

Quadro 1 - Requisitos de semelhança e diferença para harmonização de termos e conceitos	95
Quadro 2 - Exemplos de substituição de termos definidos	99
Quadro 3 - Modelos de Transformers	131
Quadro 4 - Descrição dos corpora	169
Quadro 5 - Lista de empresas e documentos selecionados para compor o Corpus da harmonização	171
Quadro 6 - Descrição da ontologia	179
Quadro 7 - Etapas da aplicação de harmonização terminológica	186
Quadro 8 - Comparativo entre as ferramentas computacionais	189
Quadro 9 - Instruções solicitadas ao LLM em língua portuguesa	204
Quadro 10 - Instrução 1 e Resposta do modelo GPT-4 32k	207
Quadro 11 - Instrução 2 e resposta do modelo GPT-4 32k	208
Quadro 12 - Instrução 3 e resposta do modelo GPT-4 32k	209
Quadro 13 - Instrução 4 e resposta do modelo GPT-4 32k	210
Quadro 14 - Instruções solicitadas ao LLM em língua inglesa	211
Quadro 15 - Resposta do GPT-4 32k em português	213
Quadro 16 - Resposta GPT-4 32k em inglês	214
Quadro 17 - Métricas do Knowledge Graph	221
Quadro 18 - Super Frame ERP	230
Quadro 19 - Frame ontológico Human_Resources	231
Quadro 20 - Frame Personnel_Management	231
Quadro 21 - Subframe Personnel_Administration	233
Quadro 22 - Conjunto de termos 1	238
Quadro 23 - Conjunto de termos 2	238
Quadro 24 - Conjunto de termos 3	238
Quadro 25 - Ficha terminológica do termo ‘empregado’	248
Quadro 26 - Ficha terminológica do termo ‘cotação’	255
Quadro 27 - Ficha terminológica do termo ‘conjunto de itens’	263
Quadro 28 - Prompts para a tarefa de harmonização de termos e conceitos.	266
Quadro 29 - Prompts desenvolvidos para o modelo GPT-4	268
Quadro 30 - Resposta prompt 1 - GPT-4 32k temperatura 0	271
Quadro 31 - Resposta prompt 1 - GPT-4 temperatura 1	271
Quadro 32 - Resultado prompt 2 - GPT-4 32k temperatura 0	273
Quadro 33 - Resultado prompt 2 - GPT-4 32k temperatura 1	274
Quadro 34 - Resultado prompt 3 - GPT-4 32k temperatura 0	275
Quadro 35 - Resultado prompt 3 - GPT-4 32k temperatura 1	275
Quadro 36 - Resultado prompt 4 - GPT-4 32k temperatura 0	277
Quadro 37 - Resultado prompt 4 - GPT-4 32k temperatura 1	278
Quadro 38 - Resultado prompt 5 - GPT-4 32k temperatura 0	279
Quadro 39 - Termos principais e variantes propostos pelo modelo GPT-4 32k	280
Quadro 40 - Resultado prompt 5 - GPT-4 32k temperatura 1	281
Quadro 41 - Resultado prompt 6 - GPT-4 temperatura 0	282
Quadro 42 - Termos principais e variantes propostos pelo modelo GPT-4	283
Quadro 43 - Resultado prompt 6 - GPT-4 temperatura 1	284



### Lista de Abreviaturas e Siglas

ERP – *Enterprise Resource Planning*

KG – *Knowledge Graph*

LC – *Linguística de Corpus*

LLM – *Large Language Model*

PLN – *Processamento de Linguagem Natural*

ISO – *International Standard Organization*

LM – *Language Model*

IA – *Inteligência Artificial*

RC – *Representação do conhecimento*

LC – *Linguística Cognitiva*

SF – *Semântica de Frames*

DiCoEnviro – *Dictionnaire Fondamental de l'environnement*

DO – *Dicionário Olímpico*

DP – *Dicionário Paralímpico*

TGT – *Teoria Geral da Terminologia*

TT – *Terminologia Textual*

ST - *Socioterminologia*

TCT – *Teoria Comunicativa da Terminologia*

TST – *Teoria Sociocognitiva da Terminologia*

MCI – *Modelo Cognitivo Idealizado*

FBT – *Frame-based Terminology*

UT – *Unidade terminológica*

## SUMÁRIO

<b><u>1 INTRODUÇÃO</u></b>	<b>18</b>
<b><u>2 LINGUAGEM E COGNIÇÃO: A SEMÂNTICA DE <i>FRAMES</i></u></b>	<b>27</b>
2.1 PRINCÍPIOS E COMPROMISSOS DA LINGUÍSTICA COGNITIVA	27
2.2 SEMÂNTICA DE <i>FRAMES</i>	36
2.2.1 SEMÂNTICA DE <i>FRAMES</i> E SUAS APLICAÇÕES	43
<b><u>3 AS BASES TERMINOLÓGICAS: TEORIAS DA TERMINOLOGIA E HARMONIZAÇÃO DE CONCEITOS E TERMOS</u></b>	<b>63</b>
3.1 TEORIAS TERMINOLÓGICAS: DA TEORIA GERAL DA TERMINOLOGIA À TERMINOLOGIA SOCIOCOGNITIVA	63
3.2 A <i>FRAME-BASED TERMINOLOGY</i>	75
3.3 HARMONIZAÇÃO DE CONCEITOS E TERMOS: DEFINIÇÕES E PRÁTICAS	82
3.3.1 HARMONIZAÇÃO DE CONCEITOS E TERMOS: METODOLOGIA E APLICAÇÃO	89
<b><u>4 A INTERFACE LINGUÍSTICO-COMPUTACIONAL: INTELIGÊNCIA ARTIFICIAL, PROCESSAMENTO DE LINGUAGEM NATURAL E <i>LARGE LANGUAGE MODEL</i></u></b>	<b>107</b>
4.1 INTELIGÊNCIA ARTIFICIAL E PROCESSAMENTO DE LINGUAGEM NATURAL: BASES, TÉCNICAS E MÉTODOS	107
4.1.1 FILOSOFIA, MATEMÁTICA, PSICOLOGIA E LINGUÍSTICA E RELAÇÕES COM IA E PLN	107
4.1.2 A HISTÓRIA DA INTELIGÊNCIA ARTIFICIAL	113
4.1.3 PROCESSAMENTO DE LINGUAGEM NATURAL	120
4.1.4 MODELOS DE LINGUAGEM E A INTELIGÊNCIA ARTIFICIAL GENERATIVA	128
4.2 REPRESENTAÇÃO DO CONHECIMENTO: ONTOLOGIAS, <i>KNOWLEDGE GRAPHS</i> E <i>FRAMES</i>	136
<b><u>5 METODOLOGIA</u></b>	<b>164</b>
5.1 LINGUÍSTICA DE <i>CORPUS</i>	164
5.1.1 PLANEJAMENTO E COMPILAÇÃO DOS <i>CORPORA</i>	168
5.2 <i>KNOWLEDGE GRAPHS</i> E <i>FRAMES</i> SEMÂNTICOS	177
5.2.1 CRIAÇÃO DO <i>KNOWLEDGE GRAPH</i> E <i>FRAMES SEMÂNTICOS</i> E HARMONIZAÇÃO DE TERMOS E CONCEITOS	181
5.2.2 FERRAMENTAS COMPUTACIONAIS PARA A CRIAÇÃO DO MAPA CONCEITUAL E DO <i>KNOWLEDGE GRAPH + FRAMES</i> SEMÂNTICOS	188
5.3 LLM: ESTRATÉGIAS DE <i>PROMPT</i>	193
5.3.1 TERMINOLOGIA E LLMs: INSTRUÇÕES E ESTRATÉGIAS DE <i>PROMPT</i> PARA A TAREFA DE HARMONIZAÇÃO TERMINOLÓGICA	201

<b>6 ANÁLISE DE DADOS</b>	<b>220</b>
<b>6.1 DO <i>KNOWLEDGE GRAPH</i> AOS <i>FRAMES</i>: REPRESENTANDO O CONHECIMENTO DO ERP</b>	<b>220</b>
<b>6.2 A METODOLOGIA SEMÂNTICO-TERMINOLÓGICA PARA HARMONIZAÇÃO DE TERMOS E CONCEITOS</b>	<b>237</b>
6.2.1 CONJUNTO DE TERMOS 1: EMPREGADO	238
6.2.2 CONJUNTO DE TERMOS 2: COTAÇÃO	248
6.2.3 CONJUNTO DE TERMOS 3: CONJUNTO DE ITENS	255
<b>6.3 LLMs E HARMONIZAÇÃO TERMINOLÓGICA</b>	<b>265</b>
6.3.2 CONJUNTO DE TERMOS: EMPREGADO	270
<b>7 CONSIDERAÇÕES FINAIS</b>	<b>290</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS</b>	<b>300</b>

## 1 INTRODUÇÃO

Com o processo de globalização e o avanço tecnológico, a produção técnica e científica acelerou a necessidade de se mais bem compreenderem as informações que eram geradas e compartilhadas, de modo a admitir que os textos técnicos apresentavam características singulares nos níveis sintático, semântico, pragmático e lexical, em especial na terminologia apresentada para estabelecer a comunicação (Barros, 2004). Além disso, tal avanço fez com que os estudos relativos à Terminologia fossem para além das universidades e alcançassem as empresas devido à necessidade de se investigarem, analisarem e descreverem as novas terminologias para, depois, aplicá-las em *softwares*.

O rápido avanço tecnológico nas Ciências da Computação, como o desenvolvimento da *World Wide Web*, impactou diretamente em diversas áreas do conhecimento, por exemplo, na Linguística Computacional, na Terminologia e nos Estudos da Tradução (Sin-Wai, 2015). Para tais disciplinas, o desenvolvimento da tecnologia modificou tanto a teoria quanto a prática. Em nível teórico, foi necessário se remodelar para entender e investigar as novas demandas na era da globalização; enquanto em nível prático, precisou-se atender às demandas de empresas para criar produtos, tais como *softwares*, que tornem o trabalho mais rápido e eficiente. Nesse sentido, o desenvolvimento de ferramentas computacionais, como as bases de dados de termos, ferramentas de gerenciamento terminológico e as *CAT Tools (Computer-assisted Translation Tools)*, precisou buscar na Computação meios para aprimorar seus recursos tecnológicos e deixar seus produtos mais acessíveis ao público.

No contexto computacional, com a transformação digital e as novas técnicas de *Machine Learning* e *Deep Learning*, criaram-se modelos de linguagem capazes de gerar novos conteúdos e exibir diversas capacidades, como, por exemplo, extrair entidades, classificar textos e gerar códigos. No âmbito da Inteligência Artificial Generativa, o uso de *Large Language Models (LLMs)* cresceu exponencialmente tanto pela possibilidade de se criarem textos altamente precisos e fluentes em diferentes estilos como pelas diversas tarefas que consegue desenvolver (Ozdemir, 2023; Kublik; Saboo, 2022). Uma das ferramentas que utiliza o LLM GPT é o ChatGPT (OpenAI, 2024). Ainda que a busca pelos modelos de linguagem tenha crescido em 2022, o primeiro modelo desenvolvido pela OpenAI foi em 2018. A título de comparação acerca da robustez dos

LLMs, o primeiro modelo, o GPT-1, utilizou um *dataset* constituído por 117 milhões de parâmetros e dados, enquanto o GPT-4, lançado em 2023, conta com 170 trilhões de parâmetros e permite realizar uma aplicação multimodal, isto, é recebe como entrada tanto textos como imagens.

Na interface linguístico-computacional, os LLMs se mostraram extremamente competentes em tarefas tradutórias. Nesse sentido, alguns estudos buscaram comparar a performance de *Large Language Models*, constituídos de trilhões de parâmetros, com modelos menores, contendo de um a sete bilhões de parâmetros, em tarefas de tradução considerando o aprendizado em contexto (Li *et al.*, 2023). Além disso, alguns pesquisadores se dedicaram a melhorar a performance de LLMs em línguas não-inglesas a partir do alinhamento semânticos entre os idiomas (Zhu *et al.*, 2023). Outros estudos focaram na compreensão de como os LLMs traduzem termos culturalmente marcados e expressões idiomáticas por meio de diferentes estratégias de *prompt* (Yao *et al.*, 2023; Raunak *et al.*, 2023). Contudo, poucos estudos focaram no desempenho de um modelo de linguagem em tarefas da Terminologia. De forma muito tímida, nota-se que o estudo terminológico atrelado aos modelos de linguagem parte de abordagens tradutórias, não se concentrando em atividades puramente terminológicas.

Considerando a era digital e as novas demandas tecnológicas relacionadas aos estudos linguísticos, esta pesquisa se insere na interface entre a Linguística Aplicada e a Computação com o propósito de buscar subsídios na Linguística Cognitiva, na Terminologia e no Processamento de Linguagem Natural (PLN) para solucionar impasses de ordem semântico-terminológico no desenvolvimento de ferramentas de gerenciamento empresarial.

Este estudo está conectado às pesquisas desenvolvidas pelo grupo de pesquisa VLHSem, *For Fostering Text Verticalization, Term Linking and Term Harmonization With Semantic Terminological approaches Project*, o qual se foca em aprimorar *softwares* de uma empresa<sup>1</sup> que já tem em suas práticas a adoção de estratégias de PLN, tradução, *Machine Learning*, *Deep Learning* e Inteligência Artificial Generativa. O projeto VLHSem conta com duas equipes de pesquisa: o time da Linguística Aplicada e o time da Computação. Essa aproximação entre as áreas surge, especialmente, porque algumas demandas exigem um olhar mais aprofundado em questões linguísticas, tradutórias e terminológicas. Em decorrência disso, percebe-se que há um crescente

---

<sup>1</sup> Conforme termo de confidencialidade e sigilo elaborado e assinado entre aluna e empresa, mantém-se em sigilo o nome da empresa ao qual este trabalho está vinculado.

interesse de empresas que buscam por modelos linguísticos para a solução de questões de ordem tecnológica, incentivando o desenvolvimento de pesquisas na área.

As empresas que desenvolvem *softwares* voltados para o gerenciamento de negócios buscam soluções para facilitar o processamento de dados e informação e a gestão entre empresas. Tradicionalmente, as instituições geravam seus dados e os armazenavam em bancos de dados separados, tornando a comunicação entre organizações e o compartilhamento de informações ineficientes. Como uma solução para tornar o funcionamento de empresas mais satisfatório e evitar colocar dados em risco, desenvolveram-se *softwares* que integrassem todas as áreas da empresa. Nesse sentido, ao compactar todas as informações e funções em um único programa, o gerenciamento empresarial se torna mais fluido, possibilitando que todos os departamentos que compõem a organização tenham acesso fácil aos dados e, conseqüentemente, há um aumento na produtividade da empresa. O *software* capaz de gerenciar uma empresa como acabamos de descrever é chamado de ERP, do inglês *Enterprise Resource Planning*. Trata-se de um programa de planejamento de recursos empresariais (ERP) que abrange todos os programas das principais áreas corporativas: compra e vendas, produção, administração de materiais, marketing, finanças e recursos humanos, fornecendo, assim, uma gestão integrada com uma visão única da empresa.

Para o desenvolvimento de um *software* desse porte, que gera e processa um grande volume de dados de diferentes áreas, é preciso trabalhar com tecnologias avançadas ligadas à Inteligência Artificial, ao *Deep Learning* e à Inteligência Artificial Generativa. Contudo, alguns aspectos, como o mapeamento dos domínios do ERP e o gerenciamento terminológico, não são supridos apenas com os pressupostos das áreas relativas à tecnologia e à computação, sendo necessário recorrer, também, a modelos linguísticos e fundamentos terminológicos que possam estruturar o *software*.

No que tange ao trabalho terminológico, com o grande volume de dados empresariais, frequentemente se emprega diferentes termos para designar o mesmo conceito, como, por exemplo, utilizar ‘empregado’, ‘colaborador’, ‘trabalhador’ e ‘operário’ para se referir a alguém que trabalha em uma empresa ou instituição. Essa prática, no entanto, torna a comunicação entre os diversos setores de uma empresa ineficiente. Para tornar a comunicação especializada mais eficiente, a *International Standard Organization* (ISO) desenvolve recomendações no contexto linguístico para lidar com demandas terminológicas. Nesse sentido, a ISO 860:2007, que trata da

harmonização de termos e conceitos, apresenta uma metodologia que permite trabalhar com os diferentes termos utilizados para se referir ao mesmo conceito.

De acordo com a ISO (2007), a recomendação 860:2007 referente à harmonização de termos e conceitos pode ser definida como uma tarefa terminológica. Seu objetivo recomendar o uso de termos e conceitos, a fim de evitar inconsistências terminológicas e tornar a comunicação mais eficiente (Pavel; Nolet, 2001; ISO, 2007). Para isso, tal documento segue os preceitos wüsterianos (Wüster, 1968, 1974; Felber, 1979) para tratar da normalização e padronização da linguagem especializada. Pela perspectiva de Wüster, na linguagem de especialidade, busca-se estabelecer um único termo para um conceito, eliminando a variação terminológica e alcançando, assim, uma comunicação unívoca. Com uma metodologia comparativa e tradicional da terminologia, a ISO 860:2007 propõe que se analise e verifique a possibilidade de harmonizar termos e conceitos de uma mesma área para que se estabeleça termo e conceito em uma relação um-a-um. Contudo, tal recomendação não considera os aspectos cognitivos e contextuais para compreender a variação terminológica e, conseqüentemente, permitir que se realize um estudo mais completo acerca da harmonização de termos e conceitos de um ou mais domínios relacionados.

Tendo em vista a lacuna contextual e cognitiva deixada pela recomendação ISO 860:2007, nesta tese, propomos uma abordagem semântico-terminológica para a harmonização de termos e conceitos que parte de duas frentes. A primeira frente condiz com o uso da Semântica de *Frames* (Fillmore, 1967, 1975, 1977, 1982, 1985) para compreender as relações entre termos e conceitos a partir da visão enciclopédica e experiencial. Tal modelo semântico para a análise linguística dos termos e conceitos parte do princípio de que não é possível compreender o significado de uma palavra (ou expressão linguística) sem acessar todo o conhecimento enciclopédico que se relaciona à palavra ou à expressão (Geeraerts, 2021).

A segunda frente se refere à *Frame-based Terminology*. Consoante à Semântica de *Frames*, a *Frame-based Terminology* (Faber, 2009, 2011, 2012, 2015; Faber; L'Homme, 2014; Faber *et al.*, 2005, 2006, 2007, 2009; Faber; Araúz, 2010; Faber; Cabezas-García, 2019) aponta que uma forma de compreender e estudar os termos no contexto especializado é a partir do comportamento da terminologia nos textos de domínios específicos. Essa aproximação entre as áreas mostra que, ao se utilizar a noção de *frames* para as análises terminológicas, torna-se evidente tanto o potencial semântico como o sintático dos termos, que incluem as descrições das relações conceptuais e as

possibilidades combinatórias da terminologia. Nesse sentido, os termos podem ser vistos como pontos de acesso para ativar o conhecimento especializado, indicando os conceitos dentro do domínio, as relações proposicionais no texto e as relações conceituais entre os conceitos e os termos (Faber, 2012).

Do ponto de vista prático, a *Frame-based Terminology* (Faber, 2012, 2014, 2015) fornece um modo de representar o conhecimento de termos e conceitos a partir da noção de *frame*. Considerando as informações das definições terminológicas, das estruturas conceituais, dos termos e demais conhecimentos prévios relacionados ao domínio específico a ser trabalhado, é possível criar uma representação externa da terminologia e suas relações conceituais por meio de uma ontologia (Faber, 2014). Uma representação do conhecimento nos moldes de uma ontologia ou de um *Knowledge Graph*, como veremos nesta tese, é fundamental para lidar com a harmonização de termos e conceitos tanto manualmente como automaticamente. Enquanto na harmonização manual a representação com os *frames* contribui para estruturar o conhecimento e situar os termos e conceitos dentro do domínio, na harmonização automática com o ChatGPT, a representação do conhecimento contribui para que o modelo de linguagem consiga harmonizar um conjunto de termos de forma mais precisa e acurada, considerando as informações do *Knowledge Graph* com *frames* semânticos.

Levando em conta o contexto de pesquisa no qual este trabalho se insere, nossa tarefa surge a partir da necessidade de dar um tratamento mais robusto às questões de harmonização terminológica que ocorrem no contexto do ERP, em língua portuguesa. Para tanto, voltamo-nos para o estudo da Semântica de *Frames* (Fillmore, 1975; 1976; 1982; 1985) com o objetivo de buscar soluções para as demandas relacionadas ao mapeamento dos domínios, bem como nos ocuparemos das questões terminológicas pelo viés da *Frame-based Terminology* (Faber *et al.* 2005; Faber *et al.* 2006; Faber *et al.* 2007; Faber; 2012, 2014, 2015) para organizar a terminologia da ferramenta, buscando a harmonização de termos e conceitos (Faulstich, 1998; Pavel; Nolet, 2001; Krieger; Finatto, 2004; Barros, 2004). Além disso, nosso trabalho também prevê o desenvolvimento de um *Knowledge Graph* com *frames* semânticos do ERP para representar o conhecimento dos setores que compõem o *software*, permitindo que se compreenda a terminologia e se realize uma harmonização mais completa. Por fim, com o desenvolvimento da nossa metodologia de base semântico-terminológica, automatizaremos a tarefa de harmonização, testando diferentes estratégias e técnicas de



*prompt* para que o *Large Language Model* consiga harmonizar termos e conceitos seguindo a nossa abordagem.

Esta tese se justifica em dois níveis. Em primeiro lugar, no nível teórico, observa-se que a Semântica de *Frames*, teoria desenvolvida por Charles Fillmore (1975; 1977; 1982; 1985), pela sua natureza enciclopédica, tem contribuído para o desenvolvimento de recursos léxico-terminológicos, como, por exemplo, dicionários e glossários, servindo como abordagem metodológica. O modelo empírico para análise semântica de Fillmore parte do princípio de que o *frame* é “[...] um sistema de conceitos relacionados de tal modo que, para entender qualquer um deles, é preciso entender toda a estrutura na qual se enquadram.” (Fillmore, 1982, p. 111). Sob essa perspectiva, a Semântica de *Frames* pode servir como princípio organizador de uma base de dados ERP, *software* que gerencia todos os departamentos de uma empresa, visto que o mapeamento dos domínios poderá ser feito considerando a aproximação entre linguagem e experiência proporcionada pelos *frames* semânticos.

Outro ponto em nível teórico corresponde à aproximação entre a Semântica de *Frames* e a Terminologia. Essa interface tem gerado bons frutos, sobretudo para o desenvolvimento de pesquisas no campo terminológico. Com o desenvolvimento da Terminologia baseada em *Frames*, *Frame-based Terminology* (Faber *et al.* 2005; Faber *et al.* 2006; Faber *et al.* 2007; Faber; 2012; 2014; 2015), pesquisas que ligam as duas disciplinas puderam buscar subsídios para investigar a Terminologia através da dimensão cognitiva. Embora as áreas da Semântica de *Frames* e da Terminologia, principalmente a Terminologia baseada em *Frames*, adotem uma visão enciclopédica da linguagem (especializada), ainda não há trabalhos terminológicos que dão conta da harmonização de termos em uma base de dados a partir da arquitetura de *frames* semânticos.

Em nível prático, percebe-se o interesse de empresas em buscar modelos linguísticos que sejam aplicáveis em *software* para suprir uma demanda tecnológica. Tendo em vista que o ERP apresenta uma terminologia diversificada e diferentes domínios, é necessário propor uma harmonização dos termos apresentados pelo ERP, assim como mapear os domínios que fazem parte desse *software* para tornar a comunicação mais eficaz e o acesso aos dados mais fácil. Nesse sentido, torna-se relevante desenvolver pesquisas na interface entre a Linguística Aplicada e a Computação, justamente para possibilitar o tratamento mais robusto da harmonização

de termos a partir de uma metodologia semântico-terminológica e aplicar tal procedimento na base de dados.

Além disso, em nível prático, a elaboração de uma representação do conhecimento arquitetada a partir de um *Knowledge Graph* com *frames* semânticos auxilia tanto na tarefa de harmonização de termos e conceitos, para que seja possível compreender os termos e sua localização no ERP, como na organização do *software* no que tange às relações conceituais e à descrição terminológicas. Mais além, considerando a transformação digital, sobretudo a criação e o uso de *Large Language Models* (LLMs), os trabalhos desenvolvidos na interface linguístico-computacional contribuem para aprimorar o desempenho de um modelo de linguagem, bem como oferecem diferentes percepções acerca das estratégias e das técnicas de *prompt* na realização de diferentes tarefas. Nesse sentido, observamos que a automatização de uma tarefa terminológica em um LLM, como a harmonização de termos e conceitos, é uma proposta inovadora, visto que poucos estudos apresentam abordagens para lidar com trabalhos atinentes aos estudos da Terminologia em *Large Language Models*.

Considerando o contexto da pesquisa e a tarefa de harmonizar termos e conceitos, este trabalho tem por objetivo, em nível teórico, investigar o papel da Semântica de *Frames* como princípio organizador de uma base de dados ERP, contribuindo para a harmonização terminológica e, em nível prático, desenvolver um *Knowledge Graph* com *frames* semânticos do ERP, bem como explorar estratégias e técnicas de *prompt* para trabalhar a harmonização de termos e conceitos utilizando o ChatGPT. No que se refere aos objetivos específicos, pretendemos (i) investigar como a aproximação entre a Semântica de *Frames* e a Terminologia baseada em *Frames* pode auxiliar no tratamento da harmonização de termos e conceitos; (ii) compreender como a visão enciclopédica assumida pela Semântica de *Frames* pode contribuir para organizar e estruturar uma base de dados ERP; (iii) apresentar uma metodologia semântico-terminológica que seja capaz de lidar com a técnica de harmonização de conceitos e termos; (iv) desenvolver um *Knowledge Graph* com *frames* semânticos do ERP para que seja possível implementar a harmonização na base de dados do *software*; e, por fim, (v) automatizar a abordagem semântico-terminológica para harmonização de termos e conceitos em um *Large Language Model*.

Após definirmos os objetivos deste trabalho, estabelecemos as seguintes premissas: (i) *frames* semânticos podem servir como princípio organizador para uma base de dados ERP (*Enterprise Resource Planning*); (ii) a arquitetura de *frames*

semânticos contribui para a garantia da harmonização terminológica em uma base de dados; (iii) uma metodologia de base semântico-terminológica é capaz de auxiliar na aplicação da harmonização de termos e conceitos; (iv) um *Knowledge Graph* com *frames* semânticos tem a capacidade de melhorar o desempenho de um LLM e fazê-lo gerar resultados mais precisos e eficazes; (v) um *Large Language Model* é capaz de realizar a harmonização de termos e conceitos com base em nossa proposta semântico-terminológica; e, finalmente, (vi) a partir de diferentes estratégias de *prompt*, um *Large Language Model* consegue entregar os resultados da harmonização similares aos feitos manualmente.

Atentando para os objetivos e as premissas deste trabalho, a Tese está organizada da seguinte forma: No capítulo 2, apresentaremos os pressupostos da Linguística Cognitiva e seus dois grandes compromissos, o compromisso de generalização e o compromisso cognitivo, para, em seguida, explorarmos a Semântica Cognitiva. Na primeira seção, trataremos da Semântica de *Frames* (Fillmore, 1975; 1977; 1982; 1985), teoria que compõe a Semântica Cognitiva, discutindo a aproximação entre linguagem e experiência. Além disso, na subseção *Semântica de Frames e suas aplicações*, abordaremos a relação entre a teoria fillmoreana e a Terminologia, apresentando dois produtos terminológicos – DiCoEnviro, EcoLexicon, Field: dicionários de expressões do futebol, Dicionário Olímpico e Dicionário Paralímpico – produzidos seguindo uma abordagem metodológica baseada em *frames*.

No capítulo 3, trataremos das questões terminológicas, traçando um retrospecto das teorias que compõem a área da Terminologia, iniciando com Wüster e a Teoria Geral da Terminologia até as linhas voltadas para a dimensão cognitiva, como a *Frame-based Terminology* (Terminologia baseada em *Frames*), foco de nossa pesquisa. Ademais, também discutiremos sobre variação terminológica, harmonização de termos e conceitos e suas definições e aplicações metodológicas em bases de dados de termos.

No capítulo 4, apresentaremos as etapas metodológicas da pesquisa. Assim, trataremos da Linguística de *Corpus*, pensando no delineamento dos *corpora* da pesquisa e nas ferramentas computacionais utilizadas para a elaboração de mapas conceituais e para a análise de dados. Em seguida, discutiremos sobre os *Knowledge Graphs* e *frames* semânticos, discorrendo sobre a elaboração dessa arquitetura e os recursos que serão utilizados para a construção da representação do conhecimento. Por fim, abordaremos as estratégias e as técnicas de *prompt* escolhidas para automatizarmos

a abordagem semântico-terminológica para a harmonização de termos e conceitos em um *Large Language Model*.

O capítulo 5 apresentará a análise de dados desta tese. Para isso, em um primeiro momento, descreveremos a construção do *Knowledge Graph* com *frames* semânticos do ERP para a tarefa de harmonização. Mais adiante, trataremos da aplicação da nossa abordagem semântico-terminológica para a harmonização de termos e conceitos de forma manual. Por último, abordaremos a automatização da nossa proposta de harmonização desenvolvida pelo LLM, considerando diferentes estratégias de *prompt* para a tarefa e relacionando com os dados obtidos manualmente.

No capítulo seis, apresentaremos as considerações finais desta tese, discorrendo sobre a importância da entrada e da participação de linguistas em pesquisas e tarefas que são, fundamentalmente, computacionais e desenvolvidas, na sua essência, por times da computação. No que cerne a nossa proposta de base semântico-terminológica para a harmonização de termos e conceitos, primeiramente, evidenciaremos as contribuições que esta pesquisa oferece ao propor uma nova visão para a tarefa de harmonização. Posteriormente, trataremos das relações entre as teorias discutidas para a elaboração dessa nova abordagem, bem como para a aplicação da harmonização de termos e conceitos tanto manualmente como automaticamente. Além disso, mostraremos as vantagens de se automatizar a tarefa de harmonizar termos e conceitos em um *Large Language Model* e como os parâmetros e as diferentes estratégias impactam nessa automatização. Por fim, apresentaremos os próximos passos e possíveis desdobramentos deste trabalho.

## 2 LINGUAGEM E COGNIÇÃO: A SEMÂNTICA DE *FRAMES*

Considerando que esta pesquisa se fundamenta fortemente na Semântica de *Frames*, este capítulo abordará os aspectos que englobam tal teoria. Para isso, inicialmente, trataremos da Linguística Cognitiva na seção 2.1, apresentando seus princípios e compromissos para, então, ocuparmo-nos da Semântica Cognitiva. Na seção 2.2, definiremos a Semântica de *Frames*, teoria desenvolvida por Charles Fillmore, e, em seguida, na seção 2.2.1, abordaremos algumas aplicações baseadas em *frames* semânticos, como o *DiCoEnviro*, o *EcoLexicon*, o *Field*: Dicionário de Expressões do Futebol, o Dicionário Olímpico e o Dicionário Paralímpico.

### 2.1 Princípios e Compromissos da Linguística Cognitiva

Nossas ideias e compreensões sobre o mundo e nós mesmos são pensadas, organizadas, estruturadas e transmitidas também por meio da linguagem. É ela que nos permite refletir e expressar nossos conhecimentos e nossas experiências, concordar ou não com os outros, falar sobre o que pensamentos e sentimentos. Como afirma Evans (2019, p. 18), “a linguagem revela a estrutura do pensamento: é uma janela para a mente”.

Inicialmente, os estudos atinentes aos aspectos da linguagem não eram considerados como científicos, visto que a Linguística não era uma ciência propriamente dita. Saussure<sup>2</sup> (2012), considerado o fundador da área, durante suas aulas de Linguística, discutia sobre as três fases que antecederam a criação da ciência da língua. Conforme explicitam os postulados saussurianos, primeiramente, intitulavam-se ‘gramática’ os estudos desenvolvidos pelos gregos e, posteriormente, pelos franceses acerca da lógica da língua, sem qualquer base científica. A segunda fase, denominada de ‘filologia’, buscava compreender e interpretar os textos, bem como o aspecto literário envolvido nesse estudo. Na terceira fase, chamada de ‘gramática comparativa’, percebeu-se que era possível comparar as línguas e, portanto, estudar as relações interlinguísticas. Embora muitos estudos que se ocupavam em descrever e compreender as linguagens tenham sido produzidos anteriormente, foi apenas com a publicação do

---

<sup>2</sup> Originalmente publicado em Saussure (1916).

Curso Geral de Linguística, obra de Saussure, mas publicado por seus alunos, que a Linguística começou, de fato, a ser vista como ciência.

Com os pressupostos saussurianos, o Estruturalismo introduziu conceitos como linguagem, língua e fala. Para essa vertente, a linguagem apresentava duas faces: a língua, que é o coletivo, e a fala, que é o individual, sendo complexo de se classificar o que pertence à sociedade e ao indivíduo (Saussure, 2012). Pensando nisso, definiu-se como objeto de estudo a língua, visto que ela podia ser modificada e era coletiva. Sob essa perspectiva, a língua era vista como um sistema autônomo, constituída por uma gramática que está presente na vida e na sociedade, elaborada no meio social, no qual a utilizamos para a comunicação. Além disso, Saussure (2012) também tratou do signo linguístico, que é composto por um significado (conceito) e um significante (imagem acústica). O primeiro – significado – se refere à representação mental do objeto em si; o segundo – significante – condiz com a união entre os sons e as respectivas letras que compõem uma palavra falada ou escrita.

Posteriormente, na década de 1950, emergiu uma nova vertente na área da linguagem denominada de Linguística Gerativa. A partir dos estudos de Noam Chomsky, mais especificamente com a publicação de seu livro “Estruturas Sintáticas”, em 1957, o gerativismo introduziu aspectos da Neurociência na Linguística. Inspirada pela Virada Cognitiva e rejeitando os postulados behavioristas, a abordagem gerativa via a linguagem como algo biológico e inato ao homem. Se, por um lado, o behaviorismo compreendia a linguagem como “um fenômeno externo ao indivíduo, um sistema de hábitos gerado como resposta a estímulos e fixado pela repetição” (Kenedy, 2012, p. 128); por outro lado, Chomsky afirmava que a linguagem não era previsível, mas sim criativa, uma vez que somos capazes de combinar diversas unidades e gerar novas frases. À luz disso, a linguagem deixou de ser o social e coletiva e passou a ser a mente humana.

Para compreender a linguagem, Chomsky (1957) apresentou dois conceitos centrais: competência e desempenho. No gerativismo, a competência se referia ao conhecimento linguístico inconsciente. Nesse sentido, a competência estava associada às regras internalizadas e à “intuição” humana de saber compor sentenças com tal conhecimento. Desse modo, o ser humano apresentava uma capacidade de gerar e entender uma quantidade infinita de sentenças. O desempenho era o uso concreto da língua. Nesse ponto, consideravam-se a criação e a compreensão das sentenças e as possibilidades de variações e impasses que poderíamos encontrar em uma situação real

de comunicação. Além disso, o desempenho se relacionava à nossa habilidade que vai além da linguagem, incluindo a atenção, a memória, a emoção, os níveis de estresse, o conhecimento de mundo, dentre outros aspectos (Kenedy, 2012).

Para a tradição gerativista, os estudos da linguagem estavam fundamentados em compromissos que priorizavam a sintaxe, atribuindo à semântica e à pragmática (Barcelona; Valenzuela, 2011) em um papel mais periférico. Para essa abordagem, a cognição era vista como modular, isto é, a sintaxe é um módulo autônomo e independente dos demais (fonologia, pragmática, semântica etc.). Além disso, a teoria chomskyana assumia que a capacidade humana de falar provém de um dispositivo inato, ou seja, o homem nasce com uma faculdade da linguagem inata para aprender outras línguas. Desse modo, acreditava-se que o conhecimento prévio e inconsciente de uma língua já estava na mente humana, e as regras direcionavam para o aprendizado e a construção de frases. Sob essa perspectiva, a linguagem não é vista (nem estudada) a partir do seu uso, mas sim das regras gramaticais (Kenedy, 2012).

Desde o surgimento da Linguística como ciência, passando pelos pressupostos estruturalistas, behavioristas e gerativistas, a linguagem foi estudada de forma descontextualizada, separando-se, no caso de Saussure, a língua da fala e, no de Chomsky, com a distinção entre competência e performance (Geeraerts, 2006). Para suprir as lacunas deixadas pela visão descontextualizada e formal, muitas teorias se preocuparam em tentar recontextualizar a linguagem, como, por exemplo, as abordagens funcionalistas. Tais vertentes se propuseram a investigar o uso e a forma da linguagem, considerando os aspectos contextuais e situacionais intrínsecos à língua. Nessa conjuntura, a Linguística Cognitiva emergiu da combinação entre diferentes tendências que eram compartilhadas, também, por estudos funcionalistas, com o objetivo de incorporar a recontextualização da linguagem em sua abordagem.

Rejeitando a visão formal da língua e a concepção das Semânticas Gerativas (Langacker, 2011), a Linguística Cognitiva (LC) surge a partir da insatisfação de estudiosos cognitivos, que discordavam do paradigma gerativo predominante, o qual focava na visão modular da linguagem, na Semântica das Condições de Verdade e na gramática gerativista (Dabrowska; Divjak, 2019). De acordo com Wen e Taylor (2021), a LC não é uma única teoria, mas uma abordagem, um movimento, uma perspectiva ou um paradigma, que abrange uma variedade de implicações advindas das Ciências Cognitivas, da Neurociência Cognitiva, da Filosofia, da Psicologia Cognitiva, da Psicologia Gestalt, da Antropologia, dos Estudos Culturais, dentre outras áreas.

Embora alguns estudos já buscassem subsídios na Ciência Cognitiva anteriormente, foi apenas no final dos anos 70 e início dos anos 80, com os trabalhos de George Lakoff, Ron Langacker e Leonard Talmy, que se passou a evidenciar a língua como um instrumento para organizar, processar e comunicar informações (Wen; Taylor, 2021; Evans; Green, 2006; Geeraerts; Cuyckens, 2007). Segundo Dabrowska e Divjak (2019), em 1989, estabeleceu-se a primeira Associação Internacional de Linguística Cognitiva (*International Cognitive Linguistics Association*) e se criou a primeira revista voltada para a área, *Cognitive Linguistics*, marcando, de fato, o surgimento da abordagem cognitiva (Geeraerts, 2006; Taylor; Littlemore, 2006). Geeraerts (2021) revela que alguns linguistas que atuaram na Semântica Gerativa se tornaram fundadores da Linguística Cognitiva, como, por exemplo, George Lakoff, Ronald Langacker e, em certa medida, Charles Fillmore, que foi uma grande inspiração para a LC, mas nunca se identificou como linguista cognitivo.

O fundamento da Linguística Cognitiva, segundo Geeraerts (2006), é a compreensão de que a linguagem é, na sua completude, sobre o significado. Sob essa perspectiva, vê-se a linguagem como um instrumento para organizar, processar e conceber informações. Enquanto Chomsky sustentava que o módulo sintático era primordial para a linguagem, a LC evidencia que a linguagem é primeiramente semântica. Para justificar essa mudança de perspectiva, Evans (2019) discorre sobre o papel da linguagem e como o significado está associado a ela. Parte-se do princípio de que a função da linguagem é expressar, codificar e externalizar nossos pensamentos e ideias, e isso ocorre pelo uso de símbolos. O significado associado a um símbolo linguístico é conectado a uma representação mental; logo, temos um conceito. Quando o significado é codificado por meio de símbolos linguísticos, eles apresentam a nossa realidade projetada, isto é, uma representação mental da realidade, construída pela mente humana e mediada pelos nossos sistemas perceptual e conceptual. Nesse sentido, nossa conceptualização, a forma como construímos o significado, considera o modo como vemos e compreendemos uma variedade de sensações, experiências, reflexos e mudanças que a nossa mente cria. Dessa forma, o que é codificado pela linguagem é construído pelo nosso sistema conceptual (um repertório de conceitos) para acessar e criar novas ideias (Evans, 2019). Assim, o significado não é apenas o reflexo do mundo, mas uma forma de modelar o mundo em que ele é corporificado a uma perspectiva de mundo (Geeraerts, 2006).



Essa forma de compreender a representação linguística a partir da estrutura conceptual é a preocupação central do campo da LC (Talmy, 2019). Conforme afirma Talmy (2019, p. 2), o objetivo da abordagem cognitiva “é verificar o sistema global integrado de estruturação conceptual na linguagem”. Desse modo, a Linguística Cognitiva investiga como as estruturas linguísticas interrelacionadas à estrutura conceptual estão organizadas e são expressas na linguagem (Evans, 2019). Para atingir esse objetivo, a Linguística Cognitiva estabelece dois grandes compromissos: o Compromisso de Generalização e o Compromisso Cognitivo (Wen; Taylor, 2021; Evans, 2019; Evans; Green, 2006). O primeiro, o Compromisso de Generalização, pode ser definido como os princípios gerais que são responsáveis por todos os aspectos da linguagem humana. Nesse sentido, afirma-se que há princípios teóricos comuns que são aplicáveis a todos os níveis da língua – nível sintático, semântico, pragmático. Logo, o compromisso da Linguística Cognitiva é identificar tais princípios, uma vez que estudiosos cognitivos não aceitam a visão modular proposta por Chomsky e passam a analisar um fenômeno linguístico considerando todos os componentes da linguagem sob princípios gerais comuns (Wen; Taylor, 2021; Evans, 2019; Evans; Green, 2006).

O Compromisso Cognitivo, por sua vez, como destacam Wen e Taylor (2021), deve considerar os achados acerca da cognição humana de outras disciplinas cognitivas, como a Inteligência Artificial, Psicologia, Neurociências, dentre outras, em seus estudos e pesquisas. Wen e Taylor (2021) afirmam que tal compromisso busca uma explicação acerca da linguagem humana, que deve estar de acordo com o que é conhecido sobre a mente e o cérebro por outras áreas. De acordo com Evans e Green (2006, p. 40-41), “[...] decorre do “Compromisso Cognitivo” que a linguagem e a organização linguística devem refletir princípios cognitivos gerais em vez de princípios cognitivos específicos da linguagem”. A título de exemplo da interdisciplinaridade entre as investigações da Linguística e outras áreas cognitivas, podemos ver a relação entre os estudos desenvolvidos por Eleanor Rosch no campo da Psicologia Cognitiva sobre categorização e sua ligação com as discussões acerca da conceptualização e do experiencialismo.

Além do compromisso de Generalização e do compromisso Cognitivo, a Linguística Cognitiva também segue princípios (ou hipóteses) basilares, que guiam a abordagem para os estudos da linguagem e mente. Wen e Taylor (2021, p. 2) destacam alguns princípios, a saber: (i) a linguagem não é autônoma, ou seja, ela faz parte da cognição; (ii) a linguagem é um “inventário” de símbolos linguísticos carregados de

significado; (iii) o significado é tudo o que compõe a linguagem; e (iv) o conhecimento linguístico surge da linguagem em uso. Vale salientar que tais hipóteses emergem, sobretudo, em contraposição a visão gerativista. Ao sustentar que a linguagem não é autônoma, a LC se opõe à visão modular da língua apresentada por Chomsky, apontando para a concepção de que não se pode investigar a língua isolando corpo e mente. Mais além, o gerativismo compreendia que o significado de uma sentença era definido a partir de condições de verdade. Entretanto, a LC prevê que a relação entre língua e mundo ocorre por meio da cognição, logo o significado está associado com a nossa experiência no mundo. Outro ponto se refere à rejeição da dicotomia competência *versus* performance, indicando que os padrões de uso da linguagem surgem a partir do conhecimento de como a língua é utilizada (Wen; Taylor, 2021; Evans, 2019; Geeraerts, 2006; Evans; Green, 2006).

Pelo viés da LC, é importante destacar alguns conceitos fundamentais para compreender a abordagem. Uma das ideias centrais da Linguística Cognitiva diz respeito ao experiencialismo corporificado. Ao rejeitar as concepções gerativistas de separar corpo e mente e da possibilidade de se analisar a língua a partir de um sistema formal, as Ciências Cognitivas passam a considerar a natureza humana e sua experiência no mundo. Nesse ínterim, a LC se inspira na Psicologia e na Filosofia, salientando a importância do experiencialismo humano e a sua natureza corpórea para investigar os fenômenos linguísticos (Wen; Taylor, 2021). Isso ocorre pelo fato de que a experiência humana, bem como o corpo humano e a sua estrutura conceptual, afetam a natureza da nossa experiência (Evans, 2019; Barcelona; Valenzuela, 2011; Evans; Green, 2006). Portanto, não se pode analisar os aspectos linguísticos isolando corpo e mente, mas é preciso considerar a corporificação do ser humano.

No que cerne à corporificação, Evans e Green (2006) apresentam uma relação entre a experiência corporificada, a cognição corporificada e o realismo experienciado. Para os autores, a noção da experiência corporificada pressupõe que temos uma visão de mundo específica do ser humano devido à natureza do seu corpo físico. Podemos observar a forma como nosso corpo afeta a nossa experiência a partir do espectro de cores, visto que nosso aparato visual possibilita identificar diversos tons, enquanto outras espécies não possuem o mesmo sistema visual e apresentam uma visão infravermelha (Evans; Green, 2006, p. 45). Como nossa natureza biológica afeta a forma como interagimos com o mundo e nossas experiências, o modo como conceptualizamos a realidade também é afetado, emergindo a noção da cognição

corporificada (Wen; Taylor, 2021; Evans, 2019). Com relação à cognição corporificada, por assumir a corporeidade como fator principal, nossa organização neurológica também é afetada pela estrutura do nosso corpo. Nesse viés, nosso corpo desempenha uma função fundamental para compreendermos e expressarmos os conceitos e a realidade (Geeraerts, 2006). Segundo Evans e Green (2006, p. 46), “só podemos falar sobre o que podemos perceber e conceber, e as coisas que podemos perceber e conceber derivam da experiência corporificada.”. Como consequência da cognição corporificada, o realismo experienciado sustenta que a língua não reflete o mundo, “[...] ela reflete a nossa construção humana única do mundo: nossa “visão de mundo”, conforme aparece para nós através das lentes da nossa corporificação” (Evans; Green, 2006, p. 48).

Dentro dos estudos que se fundamentam nos pressupostos da Linguística Cognitiva, encontramos duas subáreas: as Abordagens Cognitivas da Gramática e a Semântica Cognitiva. Tendo em vista que o significado tem um papel central no desenvolvimento de pesquisas na área da Linguística Cognitiva, para Evans e Green (2006, p. 48), “um modelo de significado (uma semântica cognitiva) deve ser delineado antes que um modelo cognitivo adequado de gramática possa ser desenvolvido”. Nesse sentido, mesmo que se sugira uma separação entre as subáreas para o desenvolvimento de pesquisas, as duas linhas de pesquisa se complementam. Com relação às Abordagens Cognitivas da Gramática, os estudos pertencentes a essa subárea investigam as unidades linguísticas simbólicas que constituem a língua (Evans; Green, 2006).

Com foco na Semântica Cognitiva (SC), há estudos que buscam investigar como as relações estabelecidas entre a experiência, a estrutura semântica e o sistema conceptual são expressas pela linguagem. Conforme afirma Geeraerts (2021), a Semântica Cognitiva (SC) se refere a um conjunto diversificado de modelos desenvolvidos dentro da Linguística Cognitiva para explorar e descrever o significado linguístico. Nesse sentido, Riemer (2010) salienta que tais abordagens são caracterizadas por uma visão holística do lugar da linguagem com a cognição. Sob o viés da SC, tais modelos incluem prototipicidade, redes radiais, metáfora, metonímia, *frames* e uma série de mecanismos atinentes às construções gramaticais.

Aprofundando-nos no âmbito da Semântica, tradicionalmente, as abordagens linguísticas tratavam do aspecto semântico dividindo o significado em (i) componente de dicionário e (ii) componente enciclopédico. Dessa forma, o componente de dicionário constituía o estudo lexical da semântica, ocupando-se do significado da palavra. O componente enciclopédico se referia ao conhecimento enciclopédico, sendo

ele externo ao conhecimento linguístico e ficando no domínio do conhecimento de mundo. Essa perspectiva se aproximava da visão formal e da hipótese da modularidade (Evans, 2019; Wen; Taylor, 2021). Considerando o contexto gerativista, os problemas semânticos encontrados nos pressupostos chomskyanos serviram de estímulo para estudiosos explorarem diferentes abordagens semânticas para além da perspectiva gerativa. Assim, as investigações atinentes à Linguística Cognitiva rejeitavam a visão de dicionário em prol da visão enciclopédica.

Pela perspectiva da Semântica Cognitiva, a linguagem serve de lente para investigar, analisar e descrever fenômenos cognitivos. De acordo com Riemer (2010), a SC rejeita a visão modular da língua, descarta as distinções entre sintaxe e semântica e entre semântica e pragmática. Nesse sentido, Evans e Green (2006) indicam que a Semântica Cognitiva é guiada por quatro princípios: (a) a estrutura conceptual é corporificada, (b) a estrutura semântica é a estrutura conceptual, (c) a representação do significado é enciclopédica, e (d) a construção do significado é conceptualização.

Com relação à estrutura conceptual ser corporificada, a Semântica Cognitiva parte da tese da corporificação, como discutido anteriormente, mostrando a importância entre a estrutura conceptual e a experiência sensorial (Riemer, 2010). Com isso, semanticistas cognitivos investigam como a natureza humana se relaciona e interage com o mundo externo e o experiencialismo. Nesse sentido, Evans e Green (2006, p. 157) afirmam que “a natureza da organização conceptual surge da experiência corporal, então parte do que torna a estrutura conceptual significativa é a experiência corporal com a qual ela está associada”.

O segundo princípio se relaciona ao aspecto de que a linguagem não se refere diretamente a entidades reais presentes no mundo externo, mas sim a estruturas conceptuais. Sob essa visão, se considerarmos que o sistema conceptual é muito maior do que o sistema linguístico, visto que também temos pensamentos, ideias e sentimentos codificados na linguagem, formamos apenas alguns conjuntos de conceitos ao associarmos o significado às palavras (Evans, 2019; Geeraerts, 2021). O terceiro princípio – a representação do significado é enciclopédica – está ligado à natureza do significado para a Semântica Cognitiva. Conforme afirmam Evans e Green (2006, p. 160), “as palavras não representam pacotes de significados bem fechados (a visão do dicionário), mas servem como “pontos de acesso” a vastos repositórios de conhecimento relacionados a um conceito ou domínio conceptual específico”. Nesse sentido, as palavras não apresentam apenas o significado convencionalizado, mas

consideram também nossa experiência e o contexto. Por fim, o quarto princípio aponta para a forma como o significado é construído no nível conceptual, isto é, a linguagem não está associada ao significado por si só, mas as unidades linguísticas são vistas como “pontos de acesso” para os conceitos (Evans; Green, 2006).

Como dito anteriormente, a Semântica Cognitiva é caracterizada por um conjunto de abordagens elaboradas no contexto da Linguística Cognitiva. Geeraerts (2021) apresenta três grupos de modelos nessa perspectiva. O primeiro grupo é composto por estudos que focam na estrutura interna da semântica em expressões da linguagem e suas relações com os vários sentidos. Fazem parte desse grupo as redes radiais (cf. Lakoff, 1987), que descrevem a estrutura de categorias em que as mais centrais irradiam para as novas instâncias, enquanto as menos centrais se estruturam a partir do centro da rede; e a teoria de protótipos (cf. Rosch, 1987), que investiga a interação entre os aspectos extensionais e intencionais de uma dada categoria.

O segundo grupo de modelos se concentra em mecanismos que dão conta da criação de novos significados, como, por exemplo, esquemas imagéticos, espaços mentais, metáfora conceptual e metonímia. A título de exemplo, a abordagem da metáfora conceptual, desenvolvida por Lakoff e Johnson (1980), reconhece que uma determinada metáfora não está restrita a um único item lexical, mas pode generalizar através de diferentes expressões (Geeraerts, 2021).

O último grupo se propõe a examinar os mecanismos da gramática das construções na linguagem. Nesse sentido, busca-se compreender o que e como os recursos gramaticais da língua contribuem para a conceptualização. Esse grupo de modelos é composto pelos estudos desenvolvidos por Langacker (1987, 1991) e Talmy (2000) acerca da gramática das construções, que analisam as relações semânticas associadas à sintaxe, à morfologia e aos mecanismos de construção. Além disso, tal grupo também conta a Semântica de *Frames*, teoria desenvolvida por Fillmore (1967, 1975, 1982, 1985), que descreve os significados de unidades lexicais que podem ser utilizadas para perspectivar uma conceptualização do mundo (Geeraerts, 2019, 2021). É válido ressaltar que, embora a Semântica de *Frames* tenha essa relação com a gramática das construções, tal abordagem também se aproxima dos estudos dos mecanismos lexicais. Em nosso trabalho, focaremos apenas a abordagem semântica de Fillmore, visto que a Semântica de *Frames* é a teoria basilar tanto para o nosso estudo semântico e para a representação do conhecimento como para as investigações terminológicas e para a aplicação da metodologia da harmonização de termos e conceitos. Assim, na

próxima seção, traçaremos um histórico da abordagem fillmoreana e seus principais conceitos de forma mais aprofundada.

## 2.2 Semântica de *Frames*

A Semântica de *Frames* (SF), teoria desenvolvida por Charles Fillmore (1967, 1975, 1977, 1982, 1985), surge no início dos anos 1970 com o propósito de evidenciar as falhas encontradas nos estudos sobre o significado e nas abordagens baseadas no formalismo e nas condições de verdade, como na Semântica das Condições de Verdade. Tal vertente da Semântica Formal, que se fundamentava em princípios lógicos e filosóficos, afirmava que era possível compreender e explicar o significado de uma sentença a partir das suas propriedades e circunstâncias (Harder, 2010). Nesse sentido, o significado era analisado com foco na linguagem e no mundo, observando se as condições são verdadeiras ou falsas em dada sentença. Contudo, esse tipo de abordagem não considerava as nuances da língua, como, por exemplo, a ambiguidade e os aspectos culturais (Harder, 2010; Evans, 2019).

Para compreendermos a Semântica de *Frames* e nos aprofundarmos nessa abordagem, é importante retomar o estudo *The Case for Case* (Fillmore, 1967), em que Fillmore explorou a estrutura de casos dos verbos em inglês. O autor parte do pressuposto de que uma sentença é composta por um verbo e por uma ou mais frases nominais, em que cada uma delas está ligada a um verbo em uma relação de caso. Considerando isso, Fillmore propôs uma lista de papéis semânticos, como agente, instrumental, dativo, factivo, locativo etc., a fim de identificar facetas gramaticalmente relevantes de um verbo (Boas, 2017). Tal conjunto de papéis semânticos era denominado de *frames* de casos, que especificam a valência semântica de um verbo. À vista disso, nessa abordagem, um quadro de casos se referia a uma sentença representada por um nó verbal e era composta por diversos elos que conectam os nós, os quais indicavam outros participantes da ação. Essa estrutura apresentava informações cruciais acerca da valência dos verbos, adjetivos e substantivos (Fillmore, 1967; Luger, 2013; Boas, 2017).

De acordo com Boas (2017) e Harder (2010), o trabalho de Fillmore sobre papéis semânticos, fundamentado apenas em uma teoria puramente linguística, apresentou muitos problemas. Como descrito por Boas (2017, p. 550), tais impasses se

referem ao “status dos papéis semânticos, incluindo problemas com a definição e a granularidade das funções semânticas, problemas que refletem generalizações de papéis cruzados e problemas de correspondência um a um de papéis”. Dada essa problemática, Fillmore abandona o estudo de casos e desenvolve um modelo em oposição aos pressupostos da Semântica das Condições de Verdade.

Em *An Alternative to Checklist Theories of Meaning*, Fillmore (1975) apresenta uma análise semântica a partir de *frames* como alternativa para as teorias fundamentadas nas condições de verdade. Nesse sentido, os *frames* deveriam ser elementos primários para a descrição e a análise do significado, e os papéis semânticos deveriam ser definidos em relação aos *frames* (Boas, 2017). Como consequência, Fillmore acredita que a teoria de protótipos, desenvolvida por Eleanor Rosch (1973, 1975), tenha uma relação com o *frame*, uma vez que a compreensão do significado a partir da perspectiva de *frames*, geralmente recorre a um protótipo. Desse modo, para Fillmore (1975), a união entre as teorias ocorre, sobretudo, pela tese de que *frames* são instâncias prototípicas do experiencialismo humano.

Vale salientar que, em seus trabalhos, Fillmore (1975, 1976, 1977) pontua o afastamento da Semântica de *Frames* das linhas teóricas mais formais, como o Estruturalismo e o Gerativismo da época, seja mostrando a relação entre protótipos e *frames* como um caminho alternativo para a análise semântica, abandonando as condições de verdade, seja tratando das questões contextuais e experienciais para compreender as cenas e *frames*. Com isso, o autor evidencia que a Semântica de *Frames* surge mais da tradição semântica empírica do que da formal. Entretanto, ela não é antagônica ao formalismo anterior, pois a teoria fillmoreana foca nas continuidades entre linguagem e experiência (Fillmore, 1982).

Somente no início dos anos 80 é que Fillmore desenvolve a primeira versão da Semântica de *Frames*, um modelo baseado na compreensão rica e necessária para a produção de um texto (Boas, 2017). No artigo *Frame Semantics* (Fillmore, 1982), Fillmore faz um retrospecto acerca do conceito de *frame*. Assim, o autor discorre sobre o uso de *frames* para fins de descrição sintática, no final da década de 50, e seus estudos referentes à descrição da valência verbal, no final dos anos 60, que marcaram o início da história da teoria. Contudo, somente em seus trabalhos posteriores que Fillmore delimita e define o que conhecemos com a Semântica de *Frames* (Boas, 2017).

Segundo Geeraerts (2021), os *frames* surgiram da ideia de que não é possível compreender o significado de uma palavra (ou expressão linguística) sem acessar todo o

conhecimento enciclopédico que se relaciona à palavra ou à expressão. Como afirma Fillmore (1982, 1985), sua proposta consiste em um sistema de conceitos, de modo que para compreender o significado de um desses conceitos, é preciso compreender todo o sistema. Assim, a Semântica de *Frames* é caracterizada como um modelo empírico para análise semântica, que segue os preceitos da Semântica Cognitiva ao adotar uma visão enciclopédica, relacionando linguagem e experiência (Petrucci, 2001).

Fillmore (1975, 1982, 1985) elabora a concepção de *frame* a partir da ideia de esquemas de F. Barlett (1932), estudo desenvolvido na área da Psicologia, e do trabalho na área da Inteligência Artificial produzido por Minsky (1974, 1975)<sup>3</sup>. No que tange à concepção de *frames* na Psicologia, tradicionalmente, compreendia-se o conceito como a unidade básica de conhecimento. Dessa forma, trabalhava-se com um modelo de representação do conhecimento a partir de uma lista de características e atributos associados ao conceito e às relações entre as entidades conceptuais (Evans, 2019). O maior obstáculo dessa perspectiva no campo da Psicologia era a dificuldade em se modelar como ocorrem as relações entre componentes de uma lista. Para dar conta desse impasse, Barlett (1932) desenvolveu a teoria de esquemas, que se tornou basilar na Psicologia Cognitiva, por possibilitar a representação do conhecimento em termos de *frames* (Evans, 2019). Como explicita Fillmore (2009), a união entre *frames* não linguísticos (advindos da Psicologia Cognitiva e da Computação) e *frames* de casos (Fillmore, 1967) permite pensar em uma semântica que vai além da valência verbal e passa a incluir estruturas amplas e multifacetadas, incorporando padrões complexos de conhecimento e terminologia.

Inspirado pelos trabalhos de Barlett (1932) e de Minsky, Fillmore aborda a distinção entre cena e *frame*. Para Fillmore (1975), a cena está relacionada a situações definidas por questões culturais, experiências humanas e estruturas institucionais no nível conceptual. O *frame*, por sua vez, refere-se ao sistema linguístico, como, por exemplo, palavras, escolhas de regras gramaticais e categorias linguísticas, assim como as escolhas linguísticas podem estar ligadas às instâncias prototípicas de cenas (Fillmore, 1975). Nessa perspectiva, o falante que aprendeu a relação entre cenas e *frames* ativa mutuamente os dois conceitos: os “*frames* estão associados na memória com outros *frames* em virtude de seu material linguístico compartilhado; e as cenas

---

<sup>3</sup> Na seção 4.2 Representação do conhecimento: Ontologias, *Knowledge Graphs* e *Frames*, abordaremos de forma mais aprofundada a concepção de *frames* de Minsky (1974, 1975).

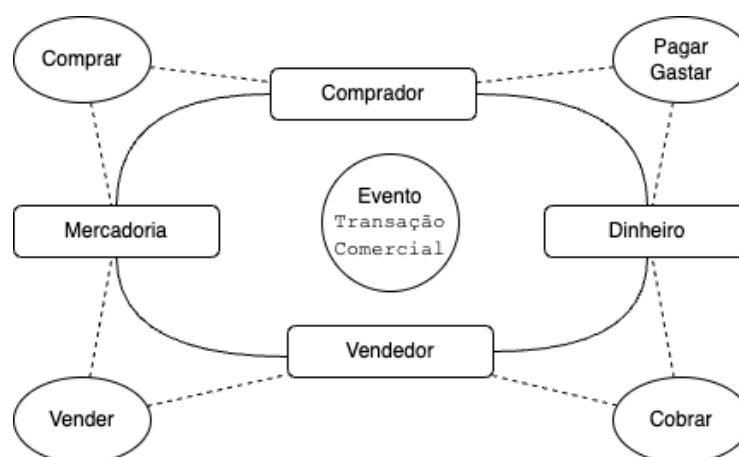


estão associadas com outras cenas em virtude da semelhança das entidades ou suas relações, ou pelo contexto de ocorrência.” (Fillmore, 1975, p. 124).

No artigo *Frame Semantics*, Fillmore (1982) abandona a distinção entre cena e *frame* e indica que este último passa a atuar tanto no nível conceptual como no nível linguístico. De acordo com Fillmore (1982, p. 117), “[a]o empregar a palavra ‘*frame*’ para indicar a maneira estruturada por meio da qual a cena é apresentada ou lembrada, podemos dizer que o *frame* estrutura os significados das palavras, e que a palavra ‘evoca’ o *frame*.”. Sob essa visão, *frame* é conceituado como uma pequena cena abstrata ou uma situação, a partir da qual, para entender o verbo em seu nível semântico, é preciso compreender todo o evento. Neste caso, sendo o *frame* uma pequena cena, a intenção é relacionar o conceito apresentado com o conhecimento prévio do interlocutor para compreender toda a situação. Além disso, os *frames* são motivados não apenas por palavras, mas também por costumes, práticas e instituições compartilhadas por uma comunidade (Fillmore, 1982, 1985; Gawron, 2019).

A fim de exemplificar o uso dos *frames*, Fillmore (1982) apresenta o *frame* Transação\_Comercial para ilustrar a ideia de que, quando acionamos um dos conceitos, automaticamente acionamos todo o nosso conhecimento e as nossas experiências, relacionando todo o sistema que envolve a transação comercial. Desse modo, ao entender uma das partes dessa situação, é possível compreender toda a situação.

Figura 1 - *Frame* Transação\_Comercial



Fonte: Elaborado pela autora.

Como podemos observar na Figura 1, o evento transação comercial envolve o comprador, o vendedor, o dinheiro, a mercadoria, bem como os verbos comprar, vender,

pagar, gastar e cobrar. Considerando isso, podemos analisar o evento a partir de diferentes perspectivas. Para Petruck (2001), ao escolhermos determinada unidade lexical para descrever a cena, assumimos diferentes perspectivas e direcionamentos do acontecimento. Assim, para compreender tal noção, podemos pensar no verbo “comprar” que foca diretamente no “comprador” e na “mercadoria”, deixando em segundo plano o “vendedor” e o “dinheiro”. O verbo “vender” foca diretamente no “vendedor” e na “mercadoria”, deixando em segundo plano o “comprador” e o “dinheiro”. A ideia do *frame*, nesse sentido, é que compreender o significado de qualquer verbo que compõe o evento transação comercial requer, até certo ponto, compreender o significado de todos eles. Portanto, a partir do *frame* e de sua estruturação, é possível compreender a natureza do evento transação comercial, pensando nos participantes e na perspectiva que envolve todas as entidades nessa atividade (Fillmore, 1977; Petruck, 2001).

A possibilidade de pensarmos e considerarmos a perspectivização é uma das vantagens que a Semântica de *Frames* oferece para os estudos lexicais. Como observado no exemplo anterior, a estrutura conceptual permite que se analise toda a cena atentando para as diferentes perspectivas. Esses diferentes pontos de vista que o *frame* apresenta nos permitem manter o princípio assumido por alguns linguistas de que os agentes interlinguísticos devem ser sujeitos. Desse modo, as diferentes perspectivas indicam que os agentes (sujeitos) e os verbos que compõem o *frame* podem coexistir, isto é, não há a exclusão de um sujeito (e verbo) em detrimento do outro, mas sim diferentes variações da situação (Gawron, 2019). No caso do *frame* *transação\_comercial*, os verbos e as demais unidades lexicais possibilitam que se compreenda toda a situação pensando tanto pela visão do comprador como do vendedor. Consoante a isso, Geeraerts (2006) salienta que uma unidade lexical pode ativar um *frame*, destacar conceitos individuais dentro do cenário e, muitas vezes, determinar uma certa perspectiva em que o *frame* é visto.

Dada a característica perspectivada do *frame*, é possível verbalizar uma determinada situação de diferentes modos (Geeraerts, 2019). Em nosso trabalho, a perspectivização, a visão enciclopédica dos *frames* e o seu aspecto contextual contribuirão para compreender o domínio do ERP e sua terminologia. No que se refere à tarefa de harmonização, essas características da SF permitem que se analise o conjunto de termos a ser harmonizado, considerando todas as perspectivas e os usos da terminologia para, então, selecionar o termo principal e as variantes. Assim, a escolha pelo termo principal

não se dará apenas pelas suas definições, mas também pelas informações extras atinentes à unidade lexical, como, por exemplo, contextos de uso, variações e exemplos. Além disso, a contextualização do *frame* e as diferentes formas de verbalizá-lo serão essenciais para o desenvolvimento dos *prompts* a serem utilizados no *Large Language Model* (LLM) (Grande Modelo de Linguagem), posto que, para a representação do conhecimento, a descrição advinda dos *frames* fornecerá informações para que o modelo possa realizar a tarefa de harmonização de forma mais precisa e eficiente.

Tendo em vista que a SF parte do princípio de que não podemos compreender o significado de uma palavra sem acessar todo o conhecimento enciclopédico associado à tal unidade lexical (Geeraerts, 2006), Gawron (2019) destaca que, para tratar do significado em um modelo conceptual como o *frame*, é necessário considerar duas propriedades acerca do significado. A primeira se refere à amplitude (abertura) dos significados das palavras, visto que a ampla variedade de significados de uma unidade se reflete na diversidade de experiências humanas. O segundo ponto condiz com a riqueza dos significados das palavras. Elaborar a definição de uma palavra não é uma tarefa simples, visto que ela pode evocar uma grande quantidade de informações extras. Considerando essas duas propriedades, observa-se que a proposição da Semântica de *Frames* (Fillmore, 1982) se baseia na conexão entre as características semântico-lexicais de amplitude e riqueza, já que a amplitude do significado depende da sua riqueza.

De acordo com Gawron (2019), o ponto de partida para a Semântica de *Frames* é assumir que as propriedades de amplitude e riqueza dos significados estão conectadas. Quando aplicamos o modelo de Fillmore para a interpretação textual, os *frames* podem ser utilizados de dois modos: (i) de forma direta, quando ligada ao sentido da palavra e (ii) a partir dos padrões entre os fatos estabelecidos no texto. Como explicita Geeraerts (2019), os *frames* permitem realizar uma descrição da situação, identificando os elementos e as entidades relevantes e o papel conceptual envolvidos naquele contexto, bem como uma descrição mais linguística, que indica como certas expressões e certos padrões gramaticais destacam as características de determinada situação. Desse modo, pode-se contrastar os conceitos presentes no texto e utilizar o conhecimento prévio na interpretação. Tal conhecimento que se aciona durante a compreensão textual são os *frames*. Quando as palavras se conectam com o conhecimento prévio compartilhado, os *frames* passam a servir como um princípio organizador para lidar com a amplitude lexical (Gawron, 2019). A afirmação de Gawron (2019) vai ao encontro do que

pretendemos neste trabalho. É justamente esse potencial em organizar as unidades lexicais que favorece o uso de *frames* para a harmonização de termos e conceitos, visto que é necessário examinar e lidar com diferentes termos em distintos contextos para selecionar o termo principal e as variantes.

No que cerne aos procedimentos metodológicos aplicados à Semântica de *Frames*, Stefanowitsch (2011) afirma que, inicialmente, a Linguística Cognitiva seguia uma abordagem não empírica, confiando em dados introspectivos. Posteriormente, com o avanço tecnológico, passou-se a utilizar a Linguística de *Corpus* para se analisarem dados e ampliarem as descrições baseadas em *frames*. Dabrowska e Divjak (2019) pontuam que a perspectiva metodológica baseada em *corpus* fez com que as abordagens da Semântica Cognitiva buscassem recursos e ferramentas já utilizados nos estudos lexicográficos para o desenvolvimento de novos métodos no campo da Linguística Cognitiva, proporcionando aos pesquisadores fontes mais confiáveis de dados e resultados (Stefanowitsch, 2011). No contexto da Semântica de *Frames*, Geeraerts (2019) destaca que, em trabalhos posteriores, a teoria fillmoreana foi enriquecida pelo uso de *corpus* como fonte de evidência empírica para a análise com *frames*, assim como o desenvolvimento de um dicionário eletrônico com descrições baseadas em *frames*.

À medida que a teoria da Semântica de *Frames* foi se desenvolvendo, podemos observar que os pressupostos de Fillmore se relacionam com a forma como a linguagem pode ser utilizada para apresentar diferentes perspectivas de uma conceptualização do mundo (Geeraerts, 2019). Em vista disso, a SF não trata apenas da nossa visão de mundo em termos de modelos conceptuais, mas como eles podem ser verbalizados de maneiras distintas. Para Fillmore (2009), a forma como a Semântica de *Frames* se estrutura, unindo os pressupostos da Psicologia, da Inteligência Artificial e da Linguística, auxilia nas análises semânticas, possibilitando “investigar de forma mais precisa e poderosa as ligações entre a gramática, o léxico e os conteúdos da experiência” (Fillmore, 2009, p. 333).

Como destaca Geeraerts (2019), a Semântica de *Frames* (SF) é uma contribuição completamente original na área da Linguística Cognitiva. Dada a natureza empírica desse modelo de análise semântica e a visão enciclopédica adotada por Fillmore, outras áreas recorrem à SF como abordagem para fundamentar seus trabalhos, como é o caso, por exemplo, da Terminologia, Lexicografia e Computação. Em seguida, abordaremos o uso da teoria fillmoreana em diferentes aplicações.

### 2.2.1 Semântica de *Frames* e suas aplicações

A Semântica de *Frames* (SF), considerando sua natureza interdisciplinar, sobretudo pela aproximação com a Psicologia, a Inteligência Artificial (IA) e a Linguística, proporciona usos não só no campo linguístico, mas também na área computacional, como no Processamento de Linguagem Natural. Fillmore (2009) ressalta que, durante o desenvolvimento da sua abordagem, alguns projetos já indicavam a relevância de se utilizarem *frames* para representação semântica em tarefas computacionais, como, por exemplo, na resolução automática de sentido (*automatic sense resolution*), marcação semântica (*semantic tagging*), perguntas e respostas (*question answering*) e tradução automática (*machine translation*).

Essa aproximação entre Linguística e Computação emergiu, principalmente, da crença de que o computador era o espelho da mente humana. De acordo com Koch e Cunha-Lima (2009), os estudos computacionais seguiam os preceitos lógicos advindos da Matemática. Esse momento foi marcado pela ideia de que “pensar é calcular”. Nesse contexto, os estudos linguístico-computacionais se fundamentavam no formalismo e racionalismo, período denominado de paradigma simbólico e que serviu de base para os estudos atinentes à Inteligência Artificial (IA). Na década de 50, com o início da era da IA, buscou-se trabalhar em projetos que fossem capazes de entender e gerar a linguagem humana. Isso possibilitou o desenvolvimento de computadores capazes de processar muitos dados e de resolver tarefas como o ser humano. Posteriormente, a criação de modelos computacionais para descrever a mente levou à criação das Ciências Cognitivas.

A Revolução Cognitiva da época encorajou os pesquisadores de diversas áreas a elaborarem programas e máquinas que recriassem o comportamento inteligente<sup>4</sup> (Russell; Norvig, 2022). Durante esse período, a crença de que o cérebro humano era um computador foi fundamental para que se investigassem as capacidades cognitivas humanas (Koch; Cunha-Lima, 2009; Riemer, 2010). No contexto linguístico-computacional, Riemer (2010, p. 270) afirma que

[...] não existe apenas uma analogia entre o funcionamento da cognição humana e um programa de computador: os dois processos são

---

<sup>4</sup> Nas seções 4.1.1 Filosofia, Matemática, Psicologia e Linguística e relações com IA e PLN, 4.1.2 A história da Inteligência Artificial e 4.1.3 Processamento de Linguagem Natural, trataremos de forma mais aprofundada sobre os estudos, as pesquisas e os caminhos da Inteligência Artificial e suas relações com outras áreas.

fundamentalmente idênticos, uma vez que a mente humana é, antes de tudo, um sistema de processamento de informação que funciona da mesma maneira, como um computador von Neumann (programa armazenado), ou seja, executando operações formais específicas sobre sequências de símbolos. Esta visão da mente como um computador está no cerne do programa clássico de pesquisa em inteligência artificial (IA).

Considerando as criações teóricas e práticas, observou-se que os computadores chegavam mais perto do que qualquer outro sistema artificial que correspondesse à complexidade do cérebro humano. Tais pressupostos e desenvolvimentos da Ciência da Computação foram basilares para a revolução chomskyana. Como explicita Chomsky (1957), a linguagem poderia ser vista como operações computacionais representadas por meio de símbolos mentais (Riemer, 2010). Dessa forma, buscou-se compreender como a linguagem é processada em mente/cérebro a partir de simulações. Dadas as restrições tecnológicas dos anos 50, assumiu-se que os modelos computacionais funcionavam em módulos. Em paralelo a isso, Chomsky defendia o conceito de modularidade na linguagem, formada pelos módulos sintático, fonológico e semântico (Koch; Cunha-Lima, 2009), indicando que o módulo sintático da linguagem é autônomo.

Como já discutido, essa visão modular da linguagem foi abandonada posteriormente, sobretudo na Linguística Cognitiva, ao compreender que não é possível separar mente e corpo. Logo, é preciso considerar nossas experiências corpóreas nos estudos linguísticos, uma vez que elas desempenham papel fundamental nas estruturas sintáticas e semânticas da língua (Barcelona; Valenzuela, 2011). Mesmo com as divergências teóricas no âmbito linguístico, é de suma importância ressaltar que os trabalhos desenvolvidos no cognitivismo clássico foram fundamentais para aproximar a Linguística da Computação. Essa união entre áreas incentivou a interdisciplinaridade, culminando em estudos linguístico-computacionais.

As investigações no campo da Linguística Computacional voltam-se para a aplicação de diversas teorias da linguagem a fim de testá-las na prática. Para a automatização de tarefas humanas em linguagem natural, exige-se um alto nível de detalhamento dos processos linguísticos para se testarem a explicitação e a integridade de uma teoria linguística (Riemer, 2010). Dada a capacidade dos computadores em processar grandes quantidades de informação, eles se tornam peças-chave na modelação do conhecimento lexical. Para exemplificar essa relação, o projeto *WordNet*<sup>5</sup>, recurso lexical pioneiro na interface da Linguística e Computação, foi criado em 1985, pela

---

<sup>5</sup> O projeto *WordNet* está disponível em: <https://wordnet.princeton.edu/> Acesso em: 27 maio 2024.

Universidade de Princeton. O *WordNet* é uma base de dados lexical *online*, similar a um tesouro<sup>6</sup>, que apresenta as relações e informações semânticas conectadas às palavras.

Tendo em vista a possibilidade de unir teoria e prática e o caráter interdisciplinar da Linguística Cognitiva, nosso foco, nesta subseção, é abordar algumas aplicações da Semântica de *Frames* em recursos linguístico-computacionais. Para isso, apresentaremos cinco recursos lexicais que são construídos com base nos *frames* semânticos de Fillmore (1967, 1975, 1982, 1985).

No que tange à relação entre Semântica de *Frames* e estudos terminológicos, percebemos que muitas investigações e linhas de pesquisa recorreram aos pressupostos da Linguística Cognitiva para fundamentar correntes que focam no viés cognitivo da Terminologia. Conforme afirma Faber (2012), a multidimensionalidade dos termos requer novas formas de examinar seu comportamento e, nas últimas décadas, passou-se a investigar com maior atenção a dimensão cognitiva dos termos. A partir da interface entre Linguística Cognitiva e Terminologia, novas linhas teóricas surgiram, como, por exemplo, a Terminologia Sociocognitiva (Temmerman, 1997, 2000, 2001, 2006) e a Terminologia baseada em *Frames* (*Frame-based Terminology*) (Faber *et al.* 2005; Faber *et al.* 2006; Faber *et al.* 2007; Faber, 2012, 2014, 2015).

O campo de interação entre Terminologia e Semântica Lexical cresceu significativamente devido à necessidade de se investigar como o significado dos termos são representados e como se comportam no texto (Faber; L’Homme, 2014). Com o avanço tecnológico e o desenvolvimento da Linguística de *Corpus*, novas abordagens terminológicas surgiram e passaram a refletir sobre as diversas dimensões do termo, dentre elas a dimensão cognitiva, que, muitas vezes, foi deixada de lado. No entanto, ao investigar as unidades terminológicas sob o viés cognitivo, é possível compreender que o significado lexical e a estrutura da língua se dão na estrutura conceptual (Faber; L’Homme, 2014; L’Homme, 2020).

Pela perspectiva da Semântica Lexical, tradicionalmente, utilizavam-se as relações como hiponímia e meronímia para representar as relações conceptuais na Terminologia. Segundo L’Homme e Robichaud (2014, p. 186), “essa perspectiva levou ao desenvolvimento de ontologias de domínio (ou estruturas menos formais) nas quais os conceitos são ligados por meio de uma rede de relações (é-um, parte-de, causa-efeito

---

<sup>6</sup> “[...] os tesouros são mecanismos que formam uma estrutura de conceitos, ou seja, criam uma lista de termos relacionados entre si, definindo os termos mais usados, quer de forma geral, quer de forma mais específica e suas correlações” (Tristão; Fachin; Alarcon, 2004).

etc.), e os termos são rótulos linguísticos desambiguados atribuídos a esses conceitos”. Contudo, esse tipo de rede de relacionamentos não dá conta das relações semânticas não-hierárquicas, como, por exemplo, agência, causa, resultado, local (Faber; L’Homme, 2014), sendo elas também essenciais para compreender e enriquecer a representação conceptual do conhecimento especializado.

Com a necessidade de se repensarem nas relações semânticas e a dimensão cognitiva dos termos, passou-se a integrar a Semântica de *Frames* nos estudos terminológicos. A teoria fillmoreana, ao adotar uma visão enciclopédica, associando linguagem e experiência, assume que a unidade lexical é construída a partir das experiências, crenças e práticas compartilhadas por instituições culturais e sociais. Por considerar as questões extralinguísticas, que são importantes tanto para os aspectos de língua geral como para as linguagens especializadas, a interface entre Semântica de *Frames* e Terminologia tem apresentado trabalhos e ferramentas computacionais inovadoras, visto que, para a área terminológica, de acordo com L’Homme (2020, p. 50),

[...] a Semântica de *Frames* é especialmente útil para explicar os participantes envolvidos no significado dos termos (especialmente, mas não exclusivamente, termos que denotam eventos e propriedades) e para descrever o comportamento sintático dos termos. Além disso, fornece um aparato para conectar as propriedades linguísticas dos termos com representações conceptuais abstratas, ou seja, *frames*. Também oferece uma maneira interessante de capturar fenômenos semânticos complexos, como modulações de significado.

Tendo em vista que o objetivo deste trabalho é investigar o papel da Semântica de *Frames* como princípio organizador de uma base de dados terminológica, apresentaremos alguns recursos lexicais, *DiCoEnviro*, *EcoLexicon*, *Field*: Dicionário de Expressões do Futebol, Dicionário Olímpico e o Dicionário Paralímpico, baseados na teoria de Charles Fillmore, mas que utilizam construções metodológicas distintas. Tal estudo serve para compreendermos melhor como os *frames* contribuem para a construção de um recurso computacional terminológico, bem como refletirmos sobre as diferentes abordagens metodológicas utilizadas pelos pesquisadores a fim de aplicarmos em nosso trabalho.

O dicionário *DiCoEnviro*<sup>7</sup> (*Dictionnaire fondamentale de l’environnement*) é um recurso computacional especializado no domínio ambiental, desenvolvido sob a

---

<sup>7</sup> Dicionário disponível em: <http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search.cgi> Acesso em: 27 maio 2024.



coordenação de Marie Claude L'Homme, do Observatoire de linguistique Sens-Texte (OLST), em Montreal, no Canadá. A ferramenta segue os princípios teóricos e metodológicos da Teoria Lexicológica Explicativo-Combinatória (*Explanatory Combinatorial Lexicology*) e se fundamenta na metodologia desenvolvida pelo Projeto FrameNet<sup>8</sup> (cf. Ruppenhofer *et al.*, 2016). Além disso, o *DiCoEnviro* está disponível em cinco línguas: Francês, Inglês, Português, Espanhol e Chinês. Na Figura 2, podemos visualizar a tela inicial do recurso *DiCoEnviro*.

Figura 2 - Tela inicial do DiCoEnviro



Fonte: *Dictionnaire fondamental de l'environnement* (2022).

Por meio da Figura 2, podemos observar que o recurso apresenta uma lista alfabética dos termos, a versão do dicionário em francês, italiano, espanhol e chinês, a documentação e a equipe que desenvolveu o *DiCoEnviro*. Além disso, no centro da tela inicial, podemos visualizar as opções de pesquisa (modo, conteúdo, nível de precisão e caixa de busca) do dicionário.

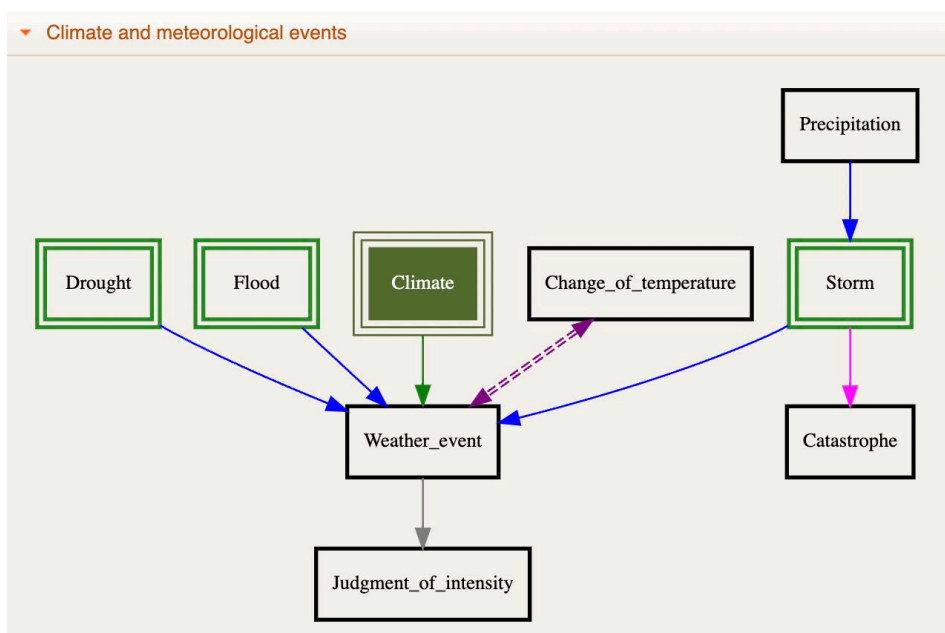
Conforme descrito na página do *DiCoEnviro*, a ideia básica do recurso lexical é direta: os significados da maioria das palavras podem ser mais bem compreendidos com base em um quadro semântico, uma descrição de um tipo de evento, relação ou entidade e seus participantes. Dessa forma, o objetivo do *DiCoEnviro* é fornecer uma descrição linguística dos termos, com foco especial nas propriedades léxico-semânticas, oferecendo um detalhamento sobre o significado linguístico do termo (estrutura

<sup>8</sup> A FrameNet é um projeto pioneiro, desenvolvido na Universidade de Berkeley, Califórnia, que tem como teoria basilar a Semântica de *Frames* (Fillmore 1976, 1977, 1982, 1985; Fillmore; Baker, 2001). A concepção que fundamenta essa base de dados parte da ideia de que os significados das palavras podem ser melhor entendidos a partir de *frames*, com uma descrição do tipo de evento ou situação, as relações e as entidades e participantes. A ferramenta está disponível em: <https://framenet.icsi.berkeley.edu/fndrupal/> Acesso em: 27 maio 2024.

actancial, combinatória e relações semânticas) e, ao mesmo tempo, evitando explicações enciclopédicas longas acerca de conceitos complexos (L'Homme; Laneville, 2010). Além dessa interface, também é possível acessar o *DiCoEnviro* na sua versão baseada em *frames*<sup>9</sup>.

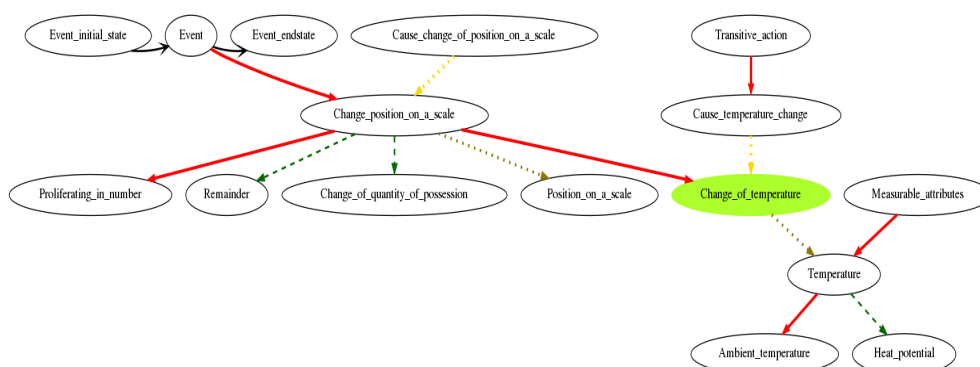
No que tange à abordagem metodológica adotada no recurso, os pesquisadores, primeiramente, recorrem a um *corpus* especializado do domínio ambiental para extração de termos e desenvolvimento de descrições terminológicas. Em seguida, os responsáveis pela parte terminológica completam as descrições a partir de dicionários e materiais disponíveis na internet. Vale mencionar que, em estudos posteriores (L'Homme; Robichaud, 2014; L'Homme, 2018, 2020), uma das etapas metodológicas na construção do *DiCoEnviro* foi buscar no Projeto FrameNet as definições de relações entre *frames* para fundamentar a ferramenta. Nesse sentido, a interface do *DiCoEnviro* é muito semelhante ao que a FrameNet apresenta em seu recurso, como podemos observar nas Figuras 3 e 4.

Figura 3 - *Frame* Evento\_climático\_e\_meteorológico



Fonte: *Frame* selecionado pela autora em *Dictionnaire fondamental de l'environnement* (2022).

<sup>9</sup> Ferramenta disponível em: <http://olst.ling.umontreal.ca/dicoenviro/framed/index.php> Acesso em: 27 maio 2024.

Figura 4 - *Frame* mudança\_de\_temperatura

Fonte: *Frame* selecionado pela autora em FrameNet (2022).

Como podemos observar nas Figuras 3 e 4, a metodologia do *DiCoEnviro* é baseada em *frames* e segue os pressupostos desenvolvidos pela FrameNet, utilizando a mesma forma de representar o conhecimento e os mesmos tipos semânticos, os relacionamentos entre *frames*, os elementos de *frames* e a sequência de eventos entre as estruturas conceituais. Com uma proposta diferente da apresentada no *DiCoEnviro*, a ferramenta *EcoLexicon* (Figura 5), desenvolvida pelo grupo de pesquisa *LexiCon*, da Universidade de Granada, na Espanha, também foca no domínio ambiental. Entretanto, o recurso se baseia na Terminologia baseada em *Frames* (Faber, 2012, 2014, 2015) – *Frame-based Terminology* – e segue apenas alguns preceitos da Semântica de *Frames* (Fillmore, 1975, 1977, 1982, 1985).

Figura 5 - Tela inicial do EcoLexicon



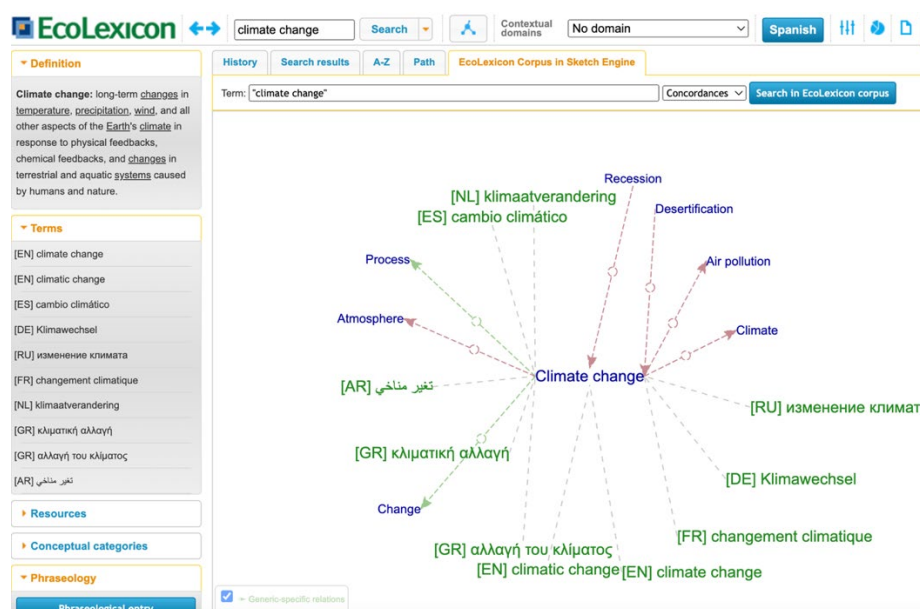
EcoLexicon is a terminological resource developed by the [LexiCon Research Group](#) at the [University of Granada](#).

EcoLexicon represents the conceptual structure of the specialized domain of the Environment in the form of a visual thesaurus. This thesaurus has been elaborated according to the theoretical premises of [Frame-Based Terminology](#). Consequently, each concept appears in the context of a specialized frame that highlights its relation to other concepts, and makes explicit its designations in English, Spanish, German, French, Russian, and Greek. Through the convergence of conceptual, linguistic and graphical information, EcoLexicon meets the needs of different users groups (people interested in the Environment, students, translators, technical writers, etc.).

Fonte: EcoLexicon (2022).

A *EcoLexicon*<sup>10</sup> é um tesouro, desenvolvido em Inglês, Espanhol, Alemão, Francês, Russo e Grego, que tem como público-alvo tradutores, estudantes, redatores e pessoas interessadas na área ambiental. Por adotar os pressupostos da Terminologia baseada em *Frames* (Faber *et al.*, 2005; Faber *et al.*, 2006; Faber *et al.*, 2007; Faber; 2012, 2014, 2015), o objetivo da ferramenta é proporcionar ao consulente uma forma dinâmica e natural do conhecimento causada, sobretudo, pelo contexto e a multidimensionalidade da base teórica aplicada. Nesse sentido, a ferramenta apresenta uma definição da unidade lexical, os equivalentes, designações, recursos conectados, fraseologias e um gráfico com as relações entre *frames*, conforme podemos observar na Figura 6.

Figura 6 - Termo 'mudança climática'



Fonte: Termo selecionado pela autora em EcoLexicon (2010).

No que se refere à abordagem metodológica adotada pelo grupo *LexiCon*, a metodologia consiste no emprego das perspectivas *bottom-up* e *top-down*. A abordagem *bottom-up* parte do uso de *corpora* multilíngues relativos ao domínio a ser investigado para a extração de informações. Com relação à abordagem *top-down*, recorre-se a dicionários especializados, materiais de referência e especialistas da área para a inclusão de mais informações (Faber, 2012). A descrição do domínio especializado é baseada em eventos, e cada domínio gera um *frame* que organiza os conceitos específicos. Assim, a estruturação dos conceitos se dá em redes verticais (hierarquias) e horizontais (não-

<sup>10</sup> O recurso EcoLexicon está disponível em: <http://ecolexicon.ugr.es/en/index.htm> Acesso em: 27 maio 2024.

hierárquicos). Para cada subdomínio que ocorre no evento, é desenvolvido um *template* que apresenta o conjunto de relações conceptuais prototípicas e as definições extraídas dos *corpora* multilíngues.

Além dos recursos lexicais apresentados, no cenário brasileiro, é necessário fazer menção a grupos de pesquisa que desenvolvem dicionários baseados em *frames*. Desde 2013, sob a coordenação da Profa. Dra. Rove Chishman, o grupo de pesquisa SemanTec, Semântica e Tecnologia, tem criado recursos lexicográficos *online* seguindo os postulados da Semântica de *Frames* e da Lexicografia Eletrônica, ou seja, parte da natureza enciclopédica e experiencial para fundamentar seus dicionários. O primeiro produto lexicográfico desenvolvido por esse grupo de pesquisa foi o *Field*: dicionário de expressões do futebol (Chishman, 2014), que tratou da linguagem do futebol. Em 2016, o grupo lançou o Dicionário Olímpico (Chishman, 2016), apresentando a riqueza do léxico das modalidades olímpicas e, em 2021, a equipe disponibilizou uma versão beta do Dicionário Paralímpico, que trata do léxico da natação paralímpica.

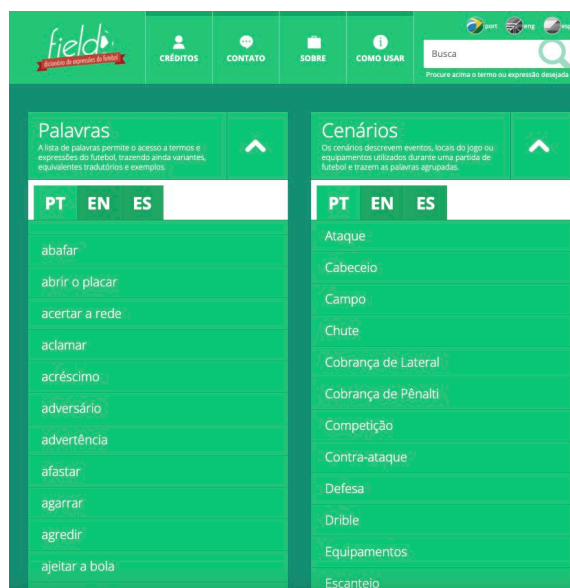
Os três dicionários desenvolvidos pelo SemanTec, como dito anteriormente, seguem os pressupostos da Semântica de *Frames* (Fillmore, 1982, 1985), partindo do conceito de *frame* para organizar e estruturar o conhecimento do domínio esportivo. Ressaltamos que os produtos lexicais desenvolvidos pelo SemanTec foram projetados para estudantes de línguas estrangeiras, pesquisadores, tradutores, linguistas e pessoas que gostam ou praticam esportes. Tendo em vista que muitas pessoas podem não conhecer a teoria de Fillmore, os conceitos teóricos foram substituídos por palavras ou expressões que os usuários possam compreender melhor a composição e estruturação dos produtos lexicográficos. Usa-se, por exemplo o termo ‘cenários’ ao invés de ‘*frames*’. Essas mudanças foram feitas justamente para facilitar a navegação “amigável” (*user-friendly*) pelos dicionários, pensando na experiência do usuário e na usabilidade.

O primeiro recurso lexical, *Field*: dicionário de expressões do futebol<sup>11</sup> (Chishman, 2014), lançado em 2014, é um dicionário trilingue (português, inglês e espanhol), que trata do léxico do futebol e é guiado pela noção de cenários (*frames*). Ao acessar o *site* do dicionário, o usuário pode buscar tanto por palavras como pelos cenários, conforme mostra a Figura 7.

---

<sup>11</sup> O recurso lexical *Field*: Dicionário de expressões do futebol está disponível em: <http://dicionariofield.com.br> Acesso em: 27 maio 2024.

Figura 7 - Tela inicial do Field: dicionário de expressões do futebol



Fonte: Chishman (2014; 2018).

Na Figura 7, podemos visualizar tanto a lista de palavras como a de cenários, além de contarmos com a possibilidade de acessar os conteúdos em diferentes línguas, em português (PT), em inglês (EN) e em espanhol (ES). Quando o usuário busca por um cenário, ele encontra a definição do cenário e as palavras relacionadas àquele *frame*, como ilustra a Figura 8.

Figura 8 - Cenário Gol

Fonte: Termo selecionado pela autora em Chishman (2014; 2018).

Como mostra a Figura 8, o cenário Gol apresenta uma definição do *frame*, explicitando as principais informações sobre aquele evento. Além disso, também encontramos as palavras (unidades lexicais) que compõem o cenário Gol e uma imagem

que representa tal cenário. Essas informações visuais e textuais contribuem para a compreensão do usuário acerca da situação, uma vez que o consulente busca em seus conhecimentos prévios o que entende sobre o evento e o que compõe aquele cenário, sendo complementado com as informações apresentadas no dicionário.

Além da possibilidade de navegar pelos cenários, o consulente também pode buscar pelas palavras. Nesse caso, ao acessar um item lexical, o usuário encontrará o termo em destaque, com a indicação da classe gramatical, o cenário em que está inserido, as variantes, os exemplos, as traduções e palavras relacionadas a ele, de acordo com o que apresenta a Figura 9.

Figura 9 - Palavra ‘Gol de Letra’

Fonte: Palavra selecionada pela autora em Chishman (2014; 2018).

Observa-se, na Figura 9, a palavra ‘gol de letra’, a classe gramatical (loc. Fig.), o cenário em que ela está inserida, um exemplo e a possibilidade de ver mais exemplos em português, equivalentes e exemplos em inglês e espanhol. É interessante destacar que, devido à natureza experiencial e enciclopédica, o recurso não apresenta as definições de cada palavra, posto que o usuário entende o significado de determinada unidade lexical a partir da contextualização presente no recurso, isto é, recorrendo aos seus conhecimentos prévios, à definição do cenário, aos exemplos e às palavras relacionadas. Com base nessas informações, o consulente é capaz de compreender o que ‘gol de letra’ significa no contexto futebolístico.

O segundo recurso lexicográfico desenvolvido pelo SemanTec foi o Dicionário Olímpico<sup>12</sup> (DO), lançado em 2016, que aborda o léxico das modalidades olímpicas. Seguindo a mesma linha teórica e metodológica do *Field*, o DO trata dos 40 esportes olímpicos, como ilustra a Figura 10.

Figura 10 - Tela inicial e modalidades do Dicionário Olímpico



Fonte: Adaptado de Chishman (2016).

Na Figura 10, podemos observar a tela inicial do Dicionário Olímpico, bem como as 40 modalidades que o compõem. É importante destacar que, considerando todas as modalidades desenvolvidas no DO, o recurso apresenta um total de 780 cenários e 3.930 palavras. No que tange à parte superior da tela inicial do dicionário, é possível encontrar os *links* para o ‘sobre’, texto que apresenta o dicionário e o histórico dos recursos desenvolvidos pelo grupo; o ‘como usar’ descreve as informações, as páginas e as diferentes formas de navegar pelo recurso; os ‘créditos’ identificam a equipe que desenvolveu o dicionário; e o ‘fale conosco’ traz a forma de entrar em contato com o grupo de pesquisa. Além disso, na tela inicial, encontramos a caixa de busca, que permite que o usuário procure pelas palavras e pelos cenários.

É possível navegar pelo DO de duas formas: i) a partir da caixa de busca, como já apresentado, em que o usuário pode pesquisar por uma palavra ou por um cenário de qualquer modalidade; ii) por meio de navegação pelas modalidades disponíveis na tela inicial. Ao acessar uma das modalidades olímpicas, o usuário encontrará uma descrição

<sup>12</sup> O Dicionário Olímpico está disponível em: <https://www.dicionarioolimpico.com.br/> Acesso em: 27 maio 2024.



geral do esporte, um mapa conceitual da modalidade, listas de cenários e palavras, curiosidades sobre determinado esporte (Você sabia?) e *links* para modalidades esportivas relacionadas, como podemos ver na Figura 11.

Figura 11 - Modalidade olímpica ‘Voleibol’

**VOLEIBOL**

O voleibol é um esporte disputado por dois times que têm como objetivo lançar uma bola por cima de uma rede e fazer com que ela atinja o chão da quadra adversária, marcando, assim, um ponto. A quadra onde o jogo ocorre mede 18 X 9m e é dividida por uma rede, de modo que cada lado meça 9 X 9m. Cada time é formado por sete jogadores e ocupa um dos lados da quadra. Cada vez que recebem a bola, os times podem tocá-la apenas três vezes. O jogo é dividido por sets e intervalos entre os sets. Vence o jogo o time que, alcançando o número de pontos necessários, conquistar três sets.

**CENÁRIOS**

**PALAVRAS**

**VOCÊ SABIA?**

Originalmente, o voleibol foi chamado de mintonette por seu criador William George Morgan. Foi em 1896, após Morgan ter apresentado o esporte em uma conferência, que um professor que assistia à apresentação sugeriu o nome *volley ball* por parecer mais adequado. O nome popular.

A jogada *china* recebeu esse nome por ter sido a marca registrada da seleção feminina chinesa do início dos anos 80.

A palavra *paragonal* foi criada no contexto do voleibol para designar um ataque que segue uma trajetória entre a paralela e a diagonal.

**VER TAMBÉM:**

Vôlei de Praia, Tênis, Badminton

Responsável(iz): Bruna da Silva  
Colaborador(es): Maximiliano Alberti (Educador Físico e técnico de voleibol)

UNISINOS, CNPq, Semantec, CAPES

Dicionário Olímpico

Tudo em direitos reservados por Flávia Reis

Quer saber mais sobre o Dicionário Olímpico? Entre em contato!

Fonte: Modalidade olímpica selecionada pela autora em Chishman (2016).

A modalidade ‘Voleibol’, como ilustra a Figura 11, apresenta uma breve descrição sobre o esporte, seu mapa conceitual acerca do funcionamento do vôlei, as unidades lexicais e os cenários. Após acessar a modalidade esportiva desejada, o usuário pode navegar tanto pela lista de palavras como pela de cenários. Quando um cenário é acessado, o consulente encontrará a descrição do cenário, as listas de palavras e cenários relacionados, um mapa conceitual do cenário e uma imagem que representa tal situação, como podemos verificar na Figura 12.

Figura 12 - Cenário Ataque

**Ataque**

O ataque é uma ação realizada próximo à rede e consiste em acertar a bola a fim de marcar um ponto. Geralmente é o terceiro contato de uma equipe com a bola. O ataque realizado pelo levantador recebe o nome de bola de segunda. O ataque realizado no primeiro toque de um time após um erro de recepção do time adversário é chamado bola de xeque. Nem todas as ações que enviam a bola para a quadra adversária são consideradas de ataque. O saque e o bloqueio não são consideradas ações de ataque. O atacante pode acertar a bola com força, realizando, assim, uma cortada, ou explorar um espaço vazio na quadra adversária, sem utilizar força, realizando uma largada ou pingada. Os ataques podem ser realizados por qualquer jogador de um time, exceto pelo jogador líbero.

**Diagrama Conceitual:**

- PONTO** (Resultado final)
- LEVANTAMENTO** (Ação anterior ao ataque)
- FUNDAMENTO** (Ação anterior ao ataque)
- ATAQUE** (Ação central)
  - Tem como objetivo marcar um
  - Vem depois de um
  - É um
- BOLA DE SEGUNDA** e **BOLA DE XEQUE** (São modos de ataque)
- CORTADA** e **LARGADA** (São tipos de ataque)

**PALAVRAS RELACIONADAS:** atacar, ataque, bola de segunda, bola de xeque, bomba, contra-ataque, cravada, cravar, explorar o bloqueio, largada.

**CENÁRIOS RELACIONADOS:** Ocorre depois de Levantamento, Pode gerar Ponto, É realizado por Jogadores, É um tipo de Jogada.

Fonte: Cenário selecionado pela autora em Chishman (2016).

É possível ver, por meio da ilustração da Figura 12, as informações que constituem o cenário Ataque. Nele, encontramos a descrição das ações que ocorrem no ‘ataque’ do voleibol, o mapa conceitual indicando a sequência de ações, uma imagem do ‘ataque’ e as listas de palavras e cenários relacionados ao ‘ataque’. Quando se acessa uma palavra do cenário, o usuário irá encontrar a unidade lexical em destaque, a classe gramatical, o equivalente e o exemplo em inglês e as palavras relacionadas àquela unidade, como ilustrado na Figura 13.

Figura 13 - Palavra ‘Bomba’

Home » Voleibol » Ataque » bomba

**bomba** *sf.*

INGLÊS: kill

EXEMPLO(S):  
Powerfull serve by Mikhaylov, but Mika gets the *kill*.

**PALAVRAS RELACIONADAS:**  
atacar, ataque, bola de segunda, bola de xeque, **bomba**, contra-ataque, cravada, cravar, explorar o bloqueio, largada

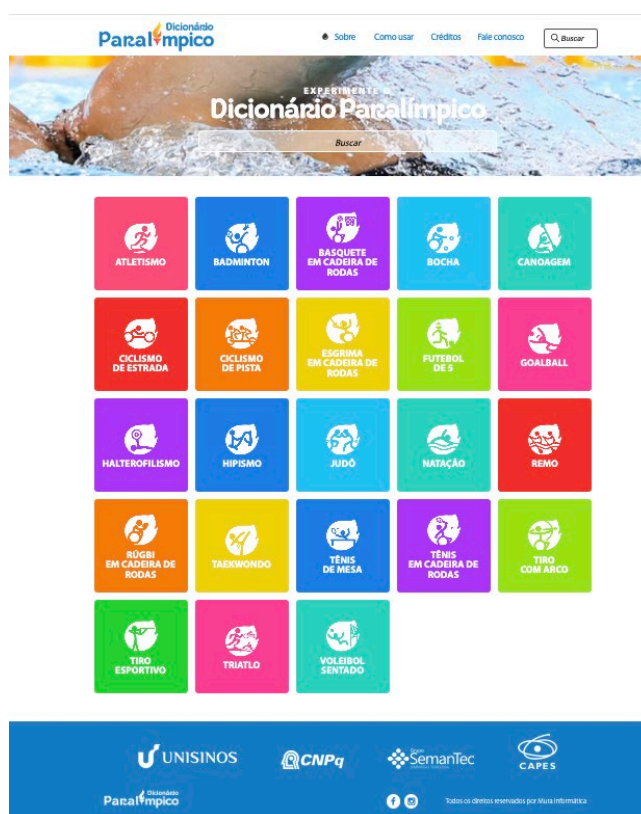
Fonte: Termo selecionado pela autora em Chishman (2016).

Na Figura 13, a palavra ‘bomba’ tem sua classe gramatical, o equivalente em inglês, ‘kill’, um exemplo, ‘Powerfull serve by Mikhaylov, but Mika gets the kill’, com a

palavra buscada em destaque (*kill* em itálico) e as palavras relacionadas, como ‘atacar’, ‘ataque’, ‘bola de segunda’, ‘bola de xeque’, ‘bomba’, ‘contra-ataque’, ‘cravada’, ‘cravar’, ‘explorar o bloqueio’ e ‘largada’. Ressalta-se que o DO não apresenta uma tradução bidirecional, isto é, o recurso não é todo traduzido. Assim, o dicionário é unidirecional: foi desenvolvido em português e apenas apresenta equivalentes e exemplos em inglês.

O último produto lexicográfico do grupo SemanTec é o Dicionário Paralímpico<sup>13</sup> (DP). Lançado em 2021, a versão beta do DP conta com a descrição da natação paralímpica. Na Figura 14, podemos visualizar a tela inicial do dicionário.

Figura 14 - Tela inicial do Dicionário Paralímpico



Fonte: Chishman (2021).

Na tela inicial do DP, conforme a Figura 14, encontramos as 23 modalidades paralímpicas. Contudo, nessa primeira versão, apenas a modalidade ‘natação’ foi desenvolvida em sua completude. Logo, não é possível acessar as demais modalidades. Quando entramos na página da ‘natação paralímpica’, de acordo com a Figura 15,

<sup>13</sup> O Dicionário Paralímpico está disponível em: <https://www.dicionarioparalimpico.com.br/> Acesso em: 27 maio 2024.

encontramos a descrição do esporte, um mapa conceitual e as listas de palavras e cenários.

Figura 15 - Modalidade: Natação paralímpica

**Home » Natação**

**NATAÇÃO**

A natação é disputada por atletas com limitações físico-motoras, deficiência visual ou deficiência intelectual.

O objetivo dos competidores é nadar determinada distância no menor tempo possível. O esporte apresenta as categorias feminina, masculina e mista.

A natação integra os Jogos Paralímpicos desde a primeira edição, realizada em Roma, na Itália, em 1960.

As primeiras medalhas brasileiras, incluindo uma de ouro, cinco de prata e uma de bronze, foram conquistadas na edição de 1984, realizada em Nova Iorque (EUA) e Stoke Mandeville (Inglaterra).

**CENÁRIOS**

- Atletas
- Brasada
- Chapada
- Classificação Funcional
- Equipamentos
- Equipos Mixtos
- Infração
- Local de competição
- Nado
- Nado borboleta
- Nado costas
- Nado crawl
- Nado livre
- Nado peito
- Oficiais de competição
- Paredada
- Prova de nado borboleta
- Prova de nado costas
- Prova de nado livre

**PALAVRAS**

- 100m borboleta
- 100m costas
- 100m livre
- 100m peito
- 150m medley
- Individual
- 200m livre
- 200m medley
- 300m medley
- 400m livre
- 50m borboleta
- 50m costas
- 50m livre
- 50m peito
- atalema
- aparelhamento do corpo
- arbitragem
- arbitragem
- aparelhador
- apito curto
- apito longo
- arbitragem
- arbitragem

**VOCÊ SABIA?**

Em Tóquio, uma das peculiaridades é que a piscina principal de 50 metros tem uma parede modular que permite transformá-la em duas piscinas separadas de 25 metros cada. A profundidade das piscinas também pode ser alterada.

Daniel Dias é o atleta com mais pódios na história do Brasil. Conquistou 24 medalhas em apenas três edições dos Jogos, sendo 14 de ouro, sete de prata e três de bronze.

**VER TAMBÉM:**

Taekwondo, Remo

UNISINOS, CNPq, Semañec, CAPES

Ajudar a melhorar a acessibilidade do site

Fonte: Modalidade selecionada pela autora em Chishman (2021).

Na Figura 15, podemos observar que a página da ‘natação paralímpica’ apresenta uma descrição do esporte, as listas de cenário e palavras, um mapa conceitual da modalidade, curiosidades (‘Você sabia?’) e esportes relacionados. Ao acessar um determinado cenário, ‘equipamentos, por exemplo, o usuário encontra uma descrição do cenário, as listas de palavras e cenários relacionados, além de uma ilustração do *frame*, como pode ser visto na Figura 16.

Figura 16 - Cenário Equipamentos

Fonte: Cenário selecionado pela autora em Chishman (2021).

A descrição do cenário Equipamentos que consta na Figura 16 inclui *hyperlinks* para acessar as palavras relacionadas ao *frame*, as listagens de palavras e cenários relacionados e uma imagem que representa tal *frame*. Além da navegação pelo cenário, o usuário também pode acessar as palavras. Nesse caso, ao buscar por uma palavra na listagem, o consulente irá encontrar a palavra em destaque, a classe gramatical, uma descrição da palavra, o cenário em que ela está inserida, as variantes, exemplos em língua portuguesa, equivalente em inglês, exemplos em língua inglesa e as palavras relacionadas àquela unidade (Figura 17).

Figura 17 - Palavra 'tapper'

Fonte: Termo selecionado pela autora em Chishman (2021).

Na Figura 17, podemos visualizar, em destaque, a palavra ‘*tapper*’ com a classe gramatical ao lado, seguida da descrição definição da palavra, o cenário em que ela está inserida, a variante, um exemplo em português, o equivalente em inglês, dois exemplos em língua inglesa e as palavras relacionadas à unidade lexical. Vale salientar que este é o primeiro produto lexical do grupo SemanTec a contar com uma descrição da palavra. Isso ocorre devido às particularidades das Paralimpíadas, uma vez que muitas pessoas não estão familiarizadas com as modalidades paralímpicas, com a linguagem inclusiva, com o atleta com deficiência ou com as diferenças entre as Olimpíadas e as Paralimpíadas.

Considerando as particularidades das Paralimpíadas, diferente dos outros dicionários do grupo, o Dicionário Paralímpico foi todo planejado de acordo com os aspectos de acessibilidade digital para pessoas com deficiência. Nesse sentido, preocupou-se em pesquisar e implementar mecanismos que auxiliem no uso do dicionário. Em razão disso, o DP apresenta recursos para baixa visão, como alto contraste, descrição das imagens, *zoom* de página e leitura de tela.

No que se refere aos fundamentos metodológicos utilizados pelo grupo de pesquisa SemanTec, os pesquisadores criaram uma metodologia diferente para a criação dos dicionários. Dessa forma, os recursos lexicais utilizam uma metodologia *Middle-out*, isto é, faz-se uso da abordagem *top-down* para a construção de um mapa conceitual e dos *frames* para, depois, aplicar a abordagem *bottom-up*, para a geração das listas de palavras e coleta de equivalentes e exemplos. Tal abordagem metodológica desenvolvida pelo grupo SemanTec foi descrita na tese de Müller (2015)<sup>14</sup>, que aproximou os *frames* semânticos e as ontologias para a representação do conhecimento do domínio jurídico. Tal trabalho também servirá como metodologia para esta tese, uma vez que nosso objetivo é representar o conhecimento do ERP no formato de *Knowledge Graph* com *frames*.

Por meio dessa explicitação das ferramentas *DiCoEnviro*, *EcoLexicon*, *Field*: dicionário de expressões do futebol, Dicionário Olímpico e Dicionário Paralímpico, trazemos evidências de que é possível desenvolver estudos na interface entre a Semântica de *Frames*, Terminologia e Lexicografia e, principalmente, aplicar a teoria fillmoreana em produtos lexicais. Além disso, ao examinarmos tais programas,

---

<sup>14</sup> Na seção 5.2.1 Criação do *Knowledge Graph* e frames semânticos e Harmonização de termos e conceitos, abordaremos de forma mais aprofundada a metodologia *Middle-out* descrita por Müller (2015).

podemos refletir sobre a metodologia desenvolvida pelos grupos de pesquisa e considerarmos quais aspectos podem ser relevantes para a nossa pesquisa, quais têm potencial para auxiliar na elaboração da metodologia semântico-terminológica voltada para o tratamento da harmonização de termos e conceitos.

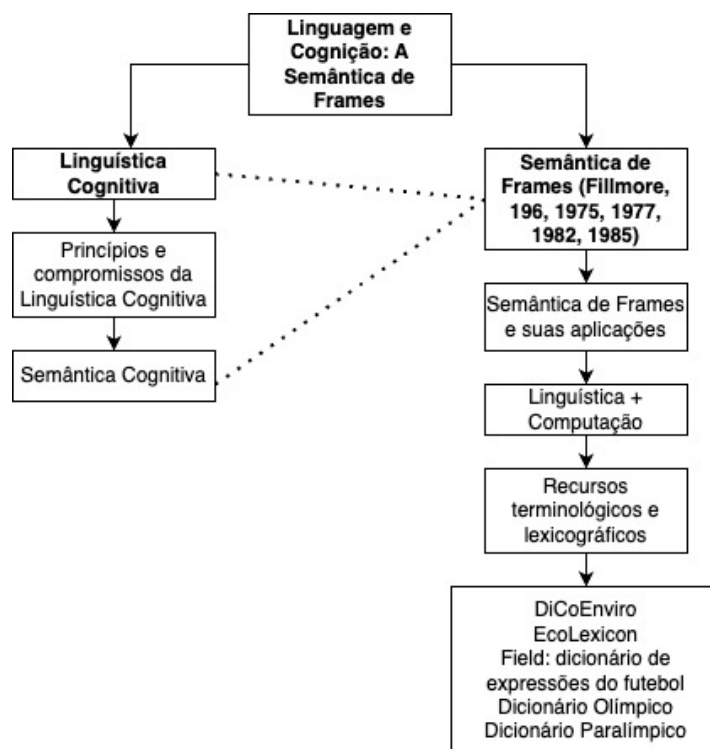
No que se refere à elaboração de um recurso lexical fundamentado no arcabouço da Linguística Cognitiva, mais especificamente na Semântica de *Frames*, podemos observar que cada grupo, a depender do objetivo da criação do seu produto lexicográfico/terminológico, desenvolve uma metodologia diferente. O *DiCoEnviro* segue a mesma metodologia desenvolvida pela FrameNet; *EcoLexicon* utilizou duas metodologias, a *bottom-up* e a *top-down*, para criar o recurso, pensando na criação dos *frames* e da lista de palavras, entre outros recursos. Em nosso trabalho, seguiremos a abordagem metodológica desenvolvida pelo grupo de pesquisa SemanTec, *Middle-out*, visto que tais procedimentos se aproximam dos nossos objetivos. A metodologia *Middle-out* une as abordagens *bottom-up* e *top-down* direcionadas à representação do conhecimento (mapas conceituais, ontologias e KGs), seleção de palavras, busca por equivalentes e exemplos e elaboração dos *frames*.

Por fim, após a apresentação das ferramentas terminológicas, vale destacar que a interface entre Semântica de *Frames* e Terminologia é um campo em expansão, justamente pelo fato de que os pressupostos da teoria desenvolvida por Fillmore são capazes de fornecer uma nova perspectiva para se analisarem as diferentes dimensões das unidades lexicais, considerando a visão enciclopédica e contextual, e se desenvolverem recursos que apresentam diferentes formas de relacionar o conteúdo linguístico com as estruturas conceituais.

Neste capítulo, nosso objetivo foi apresentar os princípios e compromissos da Linguística Cognitiva, abordagem que integra a Semântica Cognitiva para, depois, aprofundarmos nosso estudo na Semântica de *Frames*, teoria fundamental para este trabalho. Além disso, também era nosso propósito apresentar a aplicação da teoria fillmoreana em trabalhos terminológicos, como a interface utilizada para a criação dos recursos *DiCoEnviro* e *EcoLexicon*.

Considerando os pontos discutidos neste capítulo, a seguir, apresentaremos uma imagem que sintetiza o que foi trabalhado até aqui (Figura 18).

Figura 18 - Síntese do capítulo “Linguagem e Cognição: A Semântica de *Frames*”



Fonte: Elaborada pela aluna.

A Figura 18 faz uma síntese do que trabalhamos neste capítulo intitulado *Linguagem e Cognição: A Semântica de Frames*. Em um primeiro momento, discutimos sobre a Linguística Cognitiva (LC), apresentando os princípios e os compromissos da LC e a área da Semântica Cognitiva. Em seguida, aprofundamo-nos na Semântica de *Frames*, teoria desenvolvida por Charles Fillmore (1967, 1977, 1982, 1985), tratando da natureza enciclopédica e experiencial da abordagem. Além disso, discutimos acerca da conexão entre Linguística e Computação para, então, mostrarmos diferentes aplicações da Semântica de *Frames* na interface computacional. Para isso, selecionamos cinco recursos lexicográficos e terminológicos que utilizam os *frames* de Fillmore na sua concepção a fim de explorarmos diferentes metodologias.

Após esses direcionamentos, no Capítulo 3, trataremos das bases terminológicas deste trabalho, apresentando, inicialmente, as linhas teóricas da Terminologia, desde a Teoria Geral da Terminologia até a Terminologia Sociocognitiva para, então, aprofundarmos na *Frame-based Terminology* (Terminologia baseada em *Frames*), teoria que embasará as questões relacionadas à Terminologia. Por fim, abordaremos a harmonização de conceitos e termos, apresentando os processos metodológicos para a sua aplicação.



### 3 AS BASES TERMINOLÓGICAS: TEORIAS DA TERMINOLOGIA E HARMONIZAÇÃO DE CONCEITOS E TERMOS

Tendo em vista que um dos objetivos deste trabalho é tratar da harmonização de conceitos e termos em uma base de dados ERP e explorar técnicas e estratégias de *prompt* para trabalhar a harmonização utilizando o ChatGPT, neste capítulo, abordaremos a associação entre *frames* semânticos e Terminologia, bem como a harmonização terminológica. É necessário destacar que a harmonização de termos tem um papel central neste estudo, visto que esse processo poderá contribuir para tornar a comunicação mais eficaz entre as áreas que compõem o ERP e evitar inconsistências. Vale salientar também que, pela visão da Terminologia, a harmonização terminológica sofre muitas críticas pelo seu caráter normativo e por não atentar para os aspectos social, contextual e cognitivo. Considerando isso, é importante abordar essa questão a partir de uma perspectiva teórica híbrida, relacionando Semântica e Terminologia, a fim de refletir como os *frames* semânticos podem contribuir para mapear as unidades lexicais do ERP, apontar suas relações, bem como auxiliar no desenvolvimento de atividades relativas à harmonização, por exemplo, apresentar variantes, indicar novas inserções terminológicas ou confirmá-las, desaconselhar o uso de certos termos ou falsos sinônimos, dentre outras possibilidades. Na seção 3.1, abordaremos as teorias da Terminologia, apresentando desde a Teoria Geral da Terminologia até a Terminologia Sociocognitiva. Na seção 3.2, apresentaremos a *Frame-based Terminology*, uma das teorias que servirá como base para este trabalho. Na seção 3.3, definiremos a harmonização terminológica para, na subseção 3.3.1, apresentarmos a harmonização de conceitos e termos a partir da ISO 860:2007 e da NBR 13790:1997 e seus processos de aplicação.

#### 3.1 Teorias terminológicas: Da Teoria Geral da Terminologia à Terminologia Sociocognitiva

Os estudos e as práticas terminológicas vêm sendo impactados diretamente pela tecnologia e, nesse sentido, o nível teórico da Terminologia tem refletido sobre as mudanças relacionadas ao fazer terminológico. Tradicionalmente, a Terminologia examinava, analisava, descrevia e produzia glossários, dicionários e vocabulários em versão impressa. A revolução digital, por sua vez, introduziu problematizações, entre as

quais destacam-se: como aplicar os estudos terminológicos impressos em uma versão online? e, mais além, como entender o comportamento do usuário sobre a língua e tornar os produtos terminológicos mais acessíveis a seu público?. A partir desse novo contexto, a Terminologia se tornou uma disciplina multifacetada, buscando nas mais diversas áreas, como na Linguística Computacional, na Linguística de *Corpus*, na Linguística Sociocognitiva, dentre outras, suporte para conduzir seus estudos e pesquisas (Kockaert; Steurs, 2015). Pensando nisso, Geeraerts (2015, p. xix) afirma que “os desafios decorrentes dessa revolução tecnológica são consideráveis: os estudos terminológicos são estimulados a repensar seus procedimentos descritivos, programas de formação, seu formato de divulgação e, em certa medida, até mesmo seus públicos”.

No curso da história da Terminologia, observa-se que as práticas atinentes a essa área se desenvolvem desde a antiguidade. De acordo com Cabré e Lorente (2021, p. 58), “[...] a atividade terminológica é mais antiga que o desenvolvimento de um corpo teórico próprio”, visto que se organizava o conhecimento e se trabalhava com as linguagens especializadas, mas tais práticas não eram vistas como pertencentes a disciplinas científicas. Mesmo com o pleno desenvolvimento da Terminologia, foi apenas no século XVII que se reconheceu de fato a existência de vocabulários específicos de conhecimento especializado (Waquil, 2017). Como afirmam Krieger e Finatto (2004), posteriormente, no século XVIII, os estudos e as práticas terminológicas cresceram, principalmente, pela necessidade de se sistematizarem as nomenclaturas técnico-científicas dos campos da Botânica, Zoologia e Química advindas do latim e do grego.

No século XIX, com o avanço tecnológico, as práticas terminológicas tinham como meta investigar o fenômeno da internacionalização das ciências, isto é, a partir da troca de informações científicas entre diferentes países, buscava-se estabelecer uma comunicação unívoca. Nesse sentido, o objetivo era definir uma estratégia que desse conta dos padrões terminológicos e da formação de termos (Waquil, 2017; Krieger; Finatto, 2004). Tal técnica ficou conhecida como normalização terminológica.

Como explicita Waquil (2017), as primeiras discussões acerca da normalização ocorreram no Congresso de Eletricidade, em Paris, no ano de 1881. No século XX, as práticas e os estudos voltados para a normalização de termos ganharam força, especialmente com a demanda de se padronizar a terminologia da Astronomia. Como resultado, estabeleceu-se que o objetivo da normalização era garantir uma comunicação unívoca, isto é, sem qualquer ruído que pudesse interferir, gerar problemas na

comunicação entre comunidades científicas e unificar os métodos utilizados em práticas terminológicas.

Nesse contexto de investigação e desenvolvimento acerca do fenômeno da normalização terminológica, uma figura foi fundamental em relação à concepção de métodos para atingir a normalização e à criação da Terminologia como ciência. Eugen Wüster, um engenheiro austríaco, demonstrou grande interesse pelos trabalhos terminológicos. Como descreve Cabré (2003, p. 510), Wüster era “um engenheiro com forte interesse pela Ciência da Informação, outrora um esperantista convicto e um feroz defensor de uma comunicação profissional inequívoca”, tendo interesse nos estudos sobre Linguística, línguas artificiais, Filosofia, Direito, dentre outros campos (Picht, 2007).

Em 1931, Wüster defendeu sua tese de doutorado, intitulada de “A normalização internacional da linguagem a técnica especialmente na Eletrotécnica”<sup>15</sup>, que tratava da normalização na terminologia da Engenharia Elétrica. De acordo com Waquil (2017), posteriormente, o engenheiro austríaco elaborou princípios terminológicos que foram essenciais, por exemplo, para a criação da Organização Internacional de Normalização (*International Organization for Standardization – ISO*)<sup>16</sup>. Na década de 1950, Wüster lecionou na Universidade de Agricultura e Silvicultura de Viena. Nesse período, desenvolveu inúmeros trabalhos relacionados ao desenvolvimento de terminologias e glossários, resultando, em 1955, na publicação da 1ª edição de um glossário técnico-científico com mais de 10 mil entradas e, em 1959, na publicação do 2º volume. Segundo Fernández-Silva, Freixa e Faber (2012), é importante salientar que os investigadores que trabalhavam com Wüster não eram linguistas, mas especialistas de diferentes domínios científicos que estavam interessados, sobretudo, em estudar o conhecimento especializado.

No que tange aos estudos acerca da normalização terminológica, em 1968, Eugen Wüster publicou uma das suas obras mais conhecidas, *The Machine Tool. An Interlingual Dictionary of Basic Concepts* (1968), um dicionário francês-inglês de termos padronizados sistematicamente e que pretendia servir de modelo para futuros dicionários técnicos (Cabré, 2003; Hjørland, 2023). Entre os anos de 1972 e 1974,

---

<sup>15</sup> A tese de doutorado de Wüster foi escrita em alemão, intitulada “*Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik*”.

<sup>16</sup> Na seção 3.3 Harmonização de termos e conceitos: definições e práticas, abordaremos de forma mais aprofundada as normas técnicas da ISO, bem como os métodos terminológicos desenvolvidos pela Organização.

Eugen Wüster lecionou no Departamento de Linguística, onde ministrava o curso ‘Introdução à Teoria Geral de Terminologia e à Lexicografia Terminológica’ (Waquil, 2017). No ano de 1974, Wüster publicou o seu principal trabalho que objetivava estabelecer a Terminologia como disciplina. Em seu artigo intitulado ‘A Teoria Geral da Terminologia: uma área de fronteira entre Linguística, Lógica, Ontologia, Informática e as Ciências Factuais’ (Wüster, 1974), Eugen Wüster, considerando os pressupostos linguísticos da época, em especial os postulados chomskyanos que não focavam nas relações entre objetos e conceitos, tratou de incluir a ontologia e as ciências factuais em seu trabalho. Na visão de Wüster, essa inserção foi feita, pois era essencial que especialistas em linguagens e especialistas de domínio trabalhassem juntos em atividades terminológicas (Hjørland, 2023; Picht, 2007; Cabré, 2003).

Durante o curso ‘Introdução à Teoria Geral de Terminologia e à Lexicografia Terminológica’, Wüster apresentou alguns princípios que fundamentavam os estudos terminológicos. Após a sua morte, em 1977, Helmut Felber (1979), um dos discípulos de Wüster, reuniu os manuscritos e anotações do professor e publicou a obra ‘Teoria Geral da Terminologia’, no ano de 1979.

A Teoria Geral da Terminologia (TGT) tornou-se um marco na área, pois os pressupostos wüsterianos foram basilares para a Terminologia adquirir o *status* de ciência. Fernández-Silva, Freixa e Faber (2012) apontam que a Terminologia só se tornou um campo autônomo na primeira metade do século XX, em virtude da vontade de diferenciar o seu objeto de estudo, seus objetivos e métodos de trabalho dos da Linguística Geral.

Como podemos ver, Wüster se preocupou em tornar a comunicação mais eficiente entre as diferentes áreas de conhecimento e, para tanto, buscou eliminar problemas que afetassem a comunicação, como a polissemia, a ambiguidade e a homonímia, usando unidades conceituais e termos definidos com precisão (Waquil, 2017; Fernández-Silva; Freixa; Faber, 2012; Bowker, 2009). Considerando isso, Cabré (2003) afirma que Wüster pretendia atingir diversos objetivos, os quais podemos destacar dois relacionados à concepção de padronização: (i) eliminar a ambiguidade das linguagens especializadas através da padronização, tornando, assim, a comunicação mais eficaz, e (ii) comprovar os benefícios de se padronizar a Terminologia.

Tendo em mente os princípios e objetivos wüsterianos, podemos observar que a Teoria Geral da Terminologia é caracterizada pela onomasiologia, isto é, inicia-se o trabalho terminológico a partir do conceito em direção ao termo (Bowker, 2009; Cabré,

2003). De acordo com Fernández-Silva, Freixa e Faber (2012, p. 191), a TGT assume que:

- Conceitos especializados são unidades de conhecimento pertencentes ao domínio do pensamento.
- São criados através de um processo de abstração com base em propriedades compartilhadas por uma série de objetos da realidade, portanto universais e objetivos.
- São entidades bem definidas e seu conteúdo é fixado por meio de uma definição na qual são explicitadas as características essenciais.
- Os conceitos pertencem a um campo temático específico e ocupam uma posição precisa dentro do sistema conceitual desse campo. Eles mantêm relações lógicas ou ontológicas com outros conceitos dentro do campo temático.

Dessa forma, o conceito é entendido como a entidade cognitiva abstrata, a qual se refere aos objetos no mundo, e o termo é o rótulo linguístico (Faber, 2012). Como explicitam Fernández-Silva, Freixa e Faber (2012), o termo, para a TGT, é o lado formal do conceito, ou seja, ele não é uma unidade linguística, mas um símbolo que representa um conceito especializado. Por esse viés, o termo é independente do contexto. Logo, ele não mantém qualquer relação formal ou semântica com os demais elementos do discurso e está fora da influência de fatores da comunicação. À vista disso, a elaboração de um termo ocorre após a formação do conceito, visto que “os conceitos científicos são identificados por meio de rótulos, etiquetas denominativas criadas com determinadas peculiaridades que permitem fugir das ambiguidades do léxico comum” (Krieger; Finatto, 2004, p. 33). Com relação à diferença estabelecida por Wüster acerca da unidade especializada e da palavra de língua geral, percebe-se que, para a TGT, o termo apresentava uma relação direta entre o seu significado singular e a representação do conceito especializado, apresentando, assim, uma estabilidade no que se refere à forma e ao conteúdo nos textos. Justamente por essa relação entre conceito e termo é que o conceito se torna o objeto central dos estudos da Terminologia na TGT (Fernández-Silva; Freixa; Faber, 2012).

Quando se assume a normalização como objetivo principal da prática terminológica e a natureza onomasiológica da terminologia, a univocidade se torna um ponto a ser discutido e trabalhado. Segundo Waquil (2017), a comunicação unívoca é caracterizada pela ausência de ruídos, ou seja, não se aceitam termos polissêmicos, sinônimos, homônimos ou ambíguos. Bowker (2009, 2020) afirma a univocidade é tratada a partir da relação um-para-um entre conceito e termo. Desse modo, a Teoria Geral da Terminologia estabeleceu o princípio da univocidade, visto que, para se atingir

a normalização e a padronização de termos, os conceitos e termos precisariam, obrigatoriamente, ser bem definidos e, nesse sentido, eles seriam universais (todos os conceberiam e os compreenderiam da mesma forma e com a mesma definição). Logo, não se admitiam variações terminológicas e termos sinônimos nos domínios especializados.

Posteriormente, nos anos 80, a Terminologia Textual (TT) emergiu com uma visão diferente da anterior. Desenvolvida na parte germânica da Europa, a TT se fundamentava, sobretudo, nas propostas de Hartwig Kalverkämpfer e Lothar Hoffmann (Zilio, 2010). Tendo em vista os preceitos difundidos por Eugen Wüster (1968, 1974), em especial o argumento de que o trabalho terminológico deveria ser desenvolvido apenas por engenheiros e técnicos, Kalverkämpfer e Hoffmann não relacionavam seus trabalhos à Terminologia Textual, mas sim à Linguística de Linguagens Especializadas ou à Linguística do Texto Especializado (Pickbrenner; Zilio, 2015; Zilio, 2010).

Para Hoffmann (2004, p. 81), a linguagem especializada “é o conjunto de todos os recursos linguísticos que são utilizados em um âmbito comunicativo, delimitado por uma especialidade, para garantir a compreensão entre as pessoas que nela trabalham”. Dessa forma, de acordo com Hoffmann (2004), tais recursos linguísticos são encontrados no texto especializado, e a escolha e a seleção destes relacionam-se ao conteúdo especializado, à função, à finalidade comunicativa e a outros fatores que influenciam o processo comunicativo (Finatto; Kilian, 2015). Assim, essa vertente da Terminologia não investiga apenas a unidade terminológica, mas assume o texto especializado como ponto de partida para os seus estudos.

O objetivo principal da TT é compreender e descrever o funcionamento do texto internamente e externamente, de forma a analisar não apenas o termo *per se*, mas tudo o que engloba o texto e o que está conectado a ele (Zilio, 2010). Conforme afirma Leipnitz (2015), Hoffman (2005) se preocupou com a relação entre texto e gênero especializado, buscando evidenciar as características intratextuais e extratextuais, pensando na macro e na microestrutura. Atentando-nos para os aspectos sociais e comunicativos envolvidos nas análises e nos estudos de Hoffmann (2004; 2005), percebemos que o autor passa a investigar não apenas o comportamento terminológico, mas também compreende os aspectos sociais e comunicativos atrelados à Terminologia.

Para Faber e L’Homme (2022), o início da década de 1990 foi marcado como um divisor de águas no âmbito da Terminologia. Se, antes de 1990, os estudos e as práticas terminológicas se associavam à Teoria Geral da Terminologia, buscando a

normalização e padronização dos termos; após o período de 1990, novas abordagens surgiram no campo terminológico. Embora algumas teorias se conectassem mais aos pressupostos da TGT, adaptando seus métodos e suas ferramentas a essa nova era da Terminologia, outras abordagens seguiram um caminho oposto aos postulados wüsterianos. Como afirmam Faber e L’Homme (2022, p. 2),

[...] outras abordagens diferiram consideravelmente da TGT, tanto que alguns estudiosos começaram a denominá-la como Terminologia tradicional. Muitos começaram a questionar o desenvolvimento da teoria, que foi concebida para atingir o princípio da univocidade e os objetivos da normalização terminológica. Outros alegaram que o principal problema residia no fato de os princípios teóricos da TGT não estarem em consonância com as realidades linguísticas e textuais, e que a visão dos conceitos da TGT era muito inflexível.

Cabré (2003) salienta que as críticas à Teoria Geral da Terminologia emergiram de três áreas: da Ciência Cognitiva, da Comunicação e da Linguística. A Psicologia Cognitiva afirmava que era difícil separar o conhecimento especializado do conhecimento geral e que o conhecimento geral auxiliava na aquisição do conhecimento de especialidade. A Comunicação, por sua vez, trabalhava no desenvolvimento de modelos que se baseavam em diferentes cenários situacionais na comunicação geral e incluíam a especializada. Nesse sentido, a análise do discurso era primordial para se compreenderem as variações na comunicação. Na Linguística, questionava-se a rigidez com que a TGT distinguia a linguagem geral da linguagem especializada. Mais além, a Linguística passa a investigar os aspectos sociais da linguagem geral e, conseqüentemente, na especializada. Segundo Fernández-Silva, Freixa e Faber (2012), a TGT não descrevia de forma adequada o processo de categorização, a formação e as características dos conceitos. Além disso, ao assumir que os termos e conceitos são universais e bem definidos, a TGT passa a negar o caráter experiencial dos conceitos e os aspectos contextuais e culturais associados a eles.

Tendo em vista as críticas relativas à padronização terminológica e aos pressupostos da TGT, passou-se a integrar, na Terminologia, os aspectos social, comunicativo e contextual da linguagem especializada. Desse modo, quatro propostas terminológicas se destacam: a Socioterminologia, a Teoria Comunicativa da Terminologia, a Terminologia Sociocognitiva e a Terminologia baseada em *Frames*. Sendo nosso objetivo aproximar Terminologia e Semântica de *Frames*, trataremos brevemente da Socioterminologia, da Teoria Comunicativa da Terminologia e da Teoria

Sociocognitiva da Terminologia para, então, aprofundarmo-nos na Terminologia baseada em *Frames*, foco deste trabalho.

Aproximando-a dos princípios da Sociolinguística, a Socioterminologia (ST) é uma linha teórica que investiga a relação entre língua e sociedade, buscando compreender como os falantes utilizam a língua em contextos especializados. Tal vertente surgiu, principalmente, em Quebec e na França, nos anos 1970, mas realmente se estabeleceu em 1980 (Waquil, 2017). O nome Socioterminologia (*Socioterminologie*, em francês) havia sido utilizado “no início da década de 1980 por Boulanger (1981) e Lerat (1984). Slodzian (1986) também utilizou o termo em relação à Terminologia na URSS. Contudo, foi apenas com Gambier (1987) que a Socioterminologia finalmente recebeu estrutura e conteúdo” (Delavigne; Gaudin, 2022, p. 178). Segundo Delavigne e Gaudin (2022), tal abordagem foi marcada pela triangulação entre a Sociolinguística, a Terminologia e a Filosofia. Essa relação permitiu que se teorizasse acerca da Socioterminologia, a fim de contextualizar todas as pesquisas que seguiam essa vertente. Faulstich (2006) salienta que, embora o termo Socioterminologia já estivesse em uso, foi com a publicação da tese de doutorado de François Gaudin, em 1993, que se abordou o caráter social da língua na Terminologia.

Contrária aos postulados wüsterianos, que assumiam um caráter normativo e prescritivo, desconsiderando o aspecto social da língua, a Socioterminologia adotou uma abordagem descritiva da terminologia, focando o uso real da língua ao invés da abordagem prescritiva da normalização (Bowker, 2009; 2020). Dessa forma, a ST incorporou o estudo da sinonímia e da polissemia, incentivando que se fizesse um estudo mais diacrônico acerca da história da conceitualização dos nomes (e termos) (Bowker, 2020). Sob essa perspectiva, Waquil (2017) destaca que a ST se aprofundou nos aspectos sociais da linguagem especializada, atentando para a parte social e variacionista da língua.

Justamente por criticar veemente a Teoria Geral da Terminologia (Felber, 1979) e a padronização terminológica desenvolvida por Wüster (1974, 1979), Gaudin (1993) considerou a variação terminológica em seus estudos ao identificar variantes em diferentes contextos de uso (Waquil, 2017; Faber, 2012; Krieger; Finatto, 2004). Por essa perspectiva, conforme afirma Faber (2012), a abordagem socioterminológica compreendeu a padronização de termos como inconsistente, visto que a língua está em constante mudança, e as variações indicam que tanto conceitos como definições não são estáticos (Waquil, 2017). Dentre as variações identificadas, Gaudin (1993) mostrou que



a polissemia e a sinonímia são inevitáveis na terminologia, uma vez que o uso de um termo em detrimento do outro reflete o conhecimento e *status* social e profissional dos usuários, bem como as relações de poder. Em vista disso, a Socioterminologia focou as questões social e situacional da linguagem especializada, as quais podem afetar diretamente a comunicação.

Conforme afirma Delavigne e Gaudin (2022, p. 179), “o surgimento da Socioterminologia também coincidiu com o de outras teorias terminológicas”. Nesse mesmo período, outras teorias emergiram, propondo modelos alternativos, como o caso da Teoria Comunicativa da Terminologia (TCT) (Cabré, 2000, 2003) e da Teoria Sociocognitiva da Terminologia (TST) (Temmerman, 2000, 2004). Vale destacar que as abordagens terminológicas da época compartilhavam princípios teóricos. Para Delavigne e Gaudin (2022), a Socioterminologia apresentava pressupostos compatíveis com a TCT e a TST, assim como com a Terminologia Textual.

Além da Socioterminologia, a Teoria Comunicativa da Terminologia, proposta por Cabré (1993, 1999, 2000), também faz duras críticas à Teoria Geral da Terminologia (Felber, 1979), apontando para a limitação de tal perspectiva ao focar na padronização terminológica e para a dissociação entre conceito e significado (Krieger; Finatto, 2004). De acordo com Cabré e Lorente (2021, p. 59), “as principais críticas à TGT são o reducionismo, porque ela só leva em conta a função representacional da terminologia e negligencia o aspecto comunicativo”. Desse modo, nos moldes da TGT, sustentava-se que o conhecimento especializado é independente de línguas e culturas, visto que há uma universalidade de conceitos e termos.

Indo na contramão da TGT, a Teoria Comunicativa da Terminologia (TCT) fundamentava-se nos aspectos social, linguístico e cognitivo dos termos, fornecendo um arcabouço metodológico para o estudo da Terminologia (Faber, 2012; Faber; Montero-Martínez, 2019). A TCT nasceu das publicações e apresentações de Cabré, entre os anos de 1992 e 1999, propondo uma aproximação comunicativa, linguística e cognitiva da Terminologia (Cabré; Lorente, 2021, Bowker, 2020). Como descrito por Cabré e Lorente (2021, p. 59), a Teoria Comunicativa da Terminologia

não exclui a função representativa da terminologia preconizada pelo paradigma anterior, mas reivindica fundamentalmente a sua função comunicativa, tudo dentro da abordagem linguística. A TCT inova absolutamente na identificação e na delimitação do objeto científico; o termo, que é definido como uma unidade lexical, composta de significante e significado, que ativa o valor terminológico em determinados contextos comunicativos.

Pela perspectiva comunicativa da Terminologia, portanto, o termo (unidade terminológica) torna-se o objeto de estudo, como uma entidade única, o texto especializado e o conhecimento de domínio especializado (Waquil, 2017). Segundo Krieger e Finatto (2004, p. 35), a TCT postula que “[...] *a priori*, não há termos, nem palavras, mas somente unidades lexicais, tendo em vista que estas adquirem estatuto terminológico no âmbito das comunicações especializadas.”. Sob esse olhar, as unidades terminológicas são vistas como multidimensionais e se comportam como palavras de língua geral. Assim, o aspecto multidimensional das unidades especializadas está representado em três níveis: o cognitivo, o linguístico e o comunicativo. No entanto, vale salientar que, ao analisar a unidade a partir de um dos níveis, não há uma rejeição automática dos outros dois, ficando os demais níveis em segundo plano (Faber, 2012).

Outro ponto a ser destacado na Teoria Comunicativa da Terminologia é a variação terminológica. Se, por um lado, Wüster (1968, 1974) e os preceitos da TGT (Felber, 1979) buscavam eliminar a variação de termos e conceitos, visto que se pretendia atingir uma normalização terminológica; por outro lado, a TCT assumia que a variação<sup>17</sup> poderia ser vista a partir do comportamento do termo, uma vez que faz parte da linguagem natural (Waquil, 2017). Como explicitam Fernández-Silva, Freixa e Faber (2012), a abordagem de Cabré entende que a variação terminológica pode ser descrita a partir dos dois componentes do signo linguístico, o significado e o significante. Desse modo, encontramos a variação denominativa, que afeta a forma do termo, e a variação conceitual, que afeta o significado. Percebe-se, então, uma relação muito próxima entre o conceito e o termo. Isso ocorre devido à interação entre as duas variações, pois, em muitos casos, a variação denominativa está na origem ou é o resultado de alguma mudança conceitual (Fernández-Silva; Freixa; Faber, 2012).

Embora a Teoria Comunicativa da Terminologia tenha assumido as perspectivas linguística, comunicativa e cognitiva, Faber (2012) destaca que a semântica conceptual, as representações e o significado semântico não são bem descritos na TCT, necessitando de mais estudos nessa direção. Em consonância a esse pensamento, observa-se uma mudança de perspectiva nas teorias linguísticas e, como consequência, surgem

---

<sup>17</sup> Na seção 3.3, Harmonização de conceitos e termos: definições e práticas, abordaremos a variação terminológica de forma mais aprofundada, visto que ela é um fenômeno de suma importância para a tarefa de harmonização de termos e conceitos.

abordagens terminológicas de base cognitiva, como a Teoria Sociocognitiva da Terminologia e a Terminologia baseada em *Frames*.

Assim como a ST e a TCT, A Teoria Sociocognitiva da Terminologia (TST) teve críticas à Teoria Geral da Terminologia (Felber, 1979), em especial à padronização terminológica proposta por Wüster (1968, 1974), e a não inserção dos aspectos comunicativos e cognitivos nos estudos terminológicos (Temmerman, 2004). Nesse contexto, as teorias de cunho cognitivo integram pressupostos advindos da Linguística Cognitiva e da Psicologia Cognitiva e passam a estudar a estrutura de categorias e a descrição de conceitos na Terminologia (Faber, 2012). A Teoria Sociocognitiva da Terminologia, proposta por Rita Temmerman (Temmerman, 2000, 2004), foca no potencial cognitivo da linguagem especializada e na variação terminológica.

Considerando os princípios sociocognitivos, a TST não vê o termo como conceito, uma vez que parte do princípio de que não existem muitos conceitos de modo objetivo e assume a terminologia como unidade de conhecimento (Waquil, 2017). Para compreender o que é a unidade de conhecimento na perspectiva sociocognitiva da Terminologia, primeiro, precisamos entender as bases cognitivas conectadas à teoria.

Fundamentando-se nos pressupostos da Semântica Cognitiva, Temmerman (2000, 2004) afirma que a compreensão do mundo na mente humana se dá por meio de *frames* ou Modelos Cognitivos Idealizados<sup>18</sup> (MCIs) (Waquil, 2017; Krieger; Finatto, 2004; Faber, 2012). Nessa perspectiva, tais modelos conceptuais indicam que há uma relação entre eles e a unidade de conhecimento, visto que a unidade não apresenta uma estrutura prototípica e, conseqüentemente, não se comporta como conceito, mas pode ser identificada como categoria. Por esse viés, para Waquil (2017, p. 70),

uma vez que o termo passa a ser entendido mais aproximadamente como categoria e não definição, e que algumas categorias não podem ser delimitadas e traçadas rigorosamente, o mesmo “significante” que serve como termo pode se referir a conceitos e referentes em textos distintos.

---

<sup>18</sup> A teoria dos Modelos Cognitivos Idealizados (IMCs), desenvolvida por George Lakoff (1987), trata de representações mentais estáveis sobre o mundo. Esta teoria é similar aos *frames* semânticos de Fillmore (1975, 1982, 1985) pois ambas investigam as estruturas complexas de conhecimento. Os IMCs são idealizados uma vez que envolvem uma abstração das complexidades do mundo, por meio de processos perceptuais e conceptuais (Cienki, 2007). Em sua composição, encontram-se três estruturas fundamentais: a estrutura proposicional, a estrutura de esquemas imagéticos e as estruturas metafóricas e metonímicas. Essas condições são necessárias para auxiliar na nossa compreensão, organização e experiência no mundo, considerando os aspectos culturais, prototípicos e corpóreos.

À luz disso, o termo deixa de ser apenas a definição e passa a se ligar à categorização. Nesse sentido, observa-se que a unidade de conhecimento é composta por informações mais ou menos relevantes e, desse modo, aceita-se que há uma variação terminológica. Tendo em vista que a TST defende a existência da variação, os fenômenos como sinonímia e polissemia passam a ser observados nos estudos e nas análises das unidades de conhecimento.

Além dessa compreensão da conceitualização da unidade de conhecimento, Temmerman (2004) afirma que as unidades seguem a evolução e as mudanças de como compreendemos o mundo e as línguas. Isso ocorre devido à relação dos termos com os modelos cognitivos, visto que as estruturas conceptuais são transitivas e, nessa visão, assume-se que as unidades de conhecimento também sejam influenciadas por essa transição. No dizer de Temmerman (2004, p. 40),

[a] constante evolução das unidades de interpretação pode ser explicada como o resultado de vários fatores ativados simultaneamente que influenciam a classificação de um modelo cognitivo. São eles: a) a busca por um entendimento melhor e mais amplo; b) a interação entre diferentes usuários da linguagem; c) a estrutura prototípica na compreensão de categorias, que podem ser vistas simultaneamente como o resultado e como uma das causas da evolução do significado.

Com isso, podemos ver a importância que os modelos conceptuais desempenham na elaboração e na concatenação de novas ideias e novos saberes. Essa influência também se verifica na lexicalização no desenvolvimento e na criação de unidades de conhecimento (Temmerman, 2004; Waquil, 2017). Para validar seus achados, Temmerman (2004) realizou estudos terminológicos no domínio das ciências biológicas e observou que “[...] as unidades terminológicas estão em constante evolução, comportando, em consequência, sinonímia e polissemia, processo seguidamente resultante de movimentos metafóricos” (Krieger; Finatto, 2004, p. 37).

Atualmente, a Teoria Sociocognitiva da Terminologia tem se focado no estudo terminológico a partir de ontologias com o intuito de implementar as representações conceptuais. Dessa interface entre Terminologia e Ciências da Informação, surgiu a termontografia, área que busca conectar informação terminológica multilíngue às ontologias. Vale destacar que esta nova linha de estudos não aborda os Modelos Cognitivos Idealizados, conforme era trabalhado na teoria Sociocognitiva da Terminologia. De acordo com Faber (2012, p. 21), a aproximação entre termos e ontologia “não é necessariamente algo ruim desde que, a verdade seja dita,

representações linguísticas cognitivas, com a possível exceção dos *frames*, não funcionam bem em aplicativos de computador”.

Além da Teoria Sociocognitiva da Terminologia, outra abordagem também se fundamentou nos pressupostos da Linguística Cognitiva para tratar da Terminologia. A *Frame-based Terminology*, teoria desenvolvida por Pamela Faber (2009, 2011, 2012, 2015), apresenta uma linha teórica baseada em *frames* semânticos para compreender os termos. Tendo isso em mente, na próxima seção, apresentaremos a *Frame-based Terminology*, uma das teorias que embasam este trabalho.

### 3.2 A *Frame-based Terminology*

As teorias terminológicas de cunho cognitivo, como a Teoria Sociocognitiva da Terminologia (TST) e a *Frame-based Terminology* (FBT), buscam subsídios na Linguística Cognitiva (LC) para fundamentar seus estudos acerca dos termos. De acordo com Grygiel (2017), a LC pode servir como ferramenta para analisar e compreender como as linguagens especializadas são conceptualizadas e utilizadas na prática profissional. Conforme afirma Grygiel (2017, p. 8),

[...] a maior vantagem de se aplicar a Linguística Cognitiva ao estudo da linguagem especializada é colocar essa investigação não apenas no contexto de uma teoria abrangente sobre o uso da linguagem, mas também numa perspectiva ainda mais ampla, em que a linguagem faz parte da cognição e está sujeita a processos cognitivos.

Nesse sentido, como explicita Grygiel (2017), podemos perceber que o modo como a Linguística Cognitiva descreve a parte conceptual, organizando e estruturando as categorias, aproxima-se, em certa medida, dos estudos desenvolvidos pela Terminologia. Essa aproximação pode ser vista, por exemplo, nas ontologias científicas, na natureza conceptual dos termos, na composição do domínio especializado e na representação do conhecimento especializado. Ao integrarmos os modelos desenvolvidos na Linguística Cognitiva, como os Modelos Cognitivos Idealizados e os *frames*, passamos a estudar não somente o termo, mas tudo o que está atrelado à terminologia, como as estruturas conceptuais relacionadas aos termos, o contexto e os aspectos sociais e culturais compartilhados por uma comunidade.

Para Diederich (2017), dois princípios basilares da LC são cruciais para se compreenderem os termos: o significado é enciclopédico, e a construção do significado depende do contexto. Tais princípios influenciam a forma como o significado lexical é

modificado e negociado a partir do significado associativo e do contexto em que ele é utilizado. Relacionando os dois princípios com o contexto especializado, podemos observar que os significados das palavras se adaptam às atividades e aos objetivos comunicativos do domínio. Mais tarde, Diederich (2019) afirma que o significado de cada termo está sujeito a contextos experienciais e a conceptualizações do domínio como um todo. Desse modo, uma teoria terminológica que considere os aspectos cognitivos começa a dar atenção à natureza específica de cada domínio e à contextualização das unidades terminológicas, além de oferecer um caminho que relacione os termos com o seu uso real e as conceptualizações que indicam os processos no campo científico.

Pensando nessa relação entre LC e Terminologia, Pamela Faber propõe a Terminologia baseada em *Frames* (*Frame-based Terminology*) (FBT) (Faber, 2009, 2011, 2012, 2015; Faber; L’Homme, 2014; Faber *et al.*, 2005, 2006, 2007, 2009; Faber; Araújo, 2010; Faber; Cabezas-García, 2019), apontando que uma maneira de estudar a linguagem no contexto especializado é a partir do comportamento das unidades nos textos. Assumindo esse pressuposto, ao considerar a alta ocorrência de unidades terminológicas (UT) em textos de domínios específicos, Faber (2012) mostra que as unidades podem ser compreendidas como pontos de acesso para ativar o conhecimento especializado. Como consequência, “um texto rico em termos requer um conhecimento específico do domínio, os conceitos dentro dele, as relações proposicionais dentro do texto, bem como as relações conceptuais entre os conceitos do domínio” (Faber, 2012, p. 23). É sob esse olhar das relações conceptuais e da compreensão dos termos como pontos de acesso ao texto especializado que Faber aproxima sua teoria terminológica da Semântica de *Frames* (cf. Fillmore, 1976, 1977, 1982, 1985). Todavia, a Terminologia baseada em *Frames* desenvolve uma versão adaptada de alguns princípios de Fillmore.

É importante darmos um destaque quanto à ideia de Faber acerca da forma de se estudar a linguagem no contexto especializado. Tendo em vista que Faber (2009, 2011, 2012, 2015) propõe uma abordagem terminológica que se concentra na análise da terminologia no contexto especializado e no comportamento do termo dentro do texto, é possível perceber uma convergência entre essa perspectiva e os princípios defendidos por Lothar Hoffmann (2005). Na Terminologia Textual, Hoffmann não se limita à análise do termo isolado, mas considera todos os elementos que se conectam a ele. Assim, essa linha teórica incorpora tanto características intratextuais quanto extratextuais. Tais aspectos tratados por Hoffmann (2004, 2005) podem ser vistos na

forma como Faber (2012) desenvolve a *Frame-based Terminology*, colocando em foco não somente o termo, mas tudo o que engloba determinada terminologia no contexto especializado.

No que se refere à elaboração da *Frame-based Terminology*, Grygiel (2017) aponta que os *frames* são utilizados com o objetivo de estruturar o domínio especializado e criar representações de linguagem não-específicas. Dessa forma, o significado conceptual intrínseco aos textos especializados facilita a aquisição do conhecimento do domínio. Em consonância a isso, Petruck (2001, p. 4) afirma que a noção de *frame* é “[...] o tipo mais central e poderoso de estrutura de domínio que abriu caminho para uma organização do léxico baseada em *frames*, em que o *frame* fornece os fundamentos conceptuais para os sentidos relacionados de uma única palavra e palavras semanticamente relacionadas”.

Para Fillmore (1982, p. 117), “[a]o empregar a palavra ‘*frame*’ para indicar a maneira estruturada por meio da qual a cena é apresentada ou lembrada, podemos dizer que o *frame* estrutura o significado das palavras e que a palavra ‘evoca’ o *frame*”. Da mesma forma que as palavras evocam *frames*, apresentando todo o contexto e estruturas relacionadas nas quais estão inseridas, Faber (2012) afirma que as unidades terminológicas também deveriam evocar *frames* e fornecer exemplos em uso. Faber (2012) assume essa premissa, uma vez que a descrição de um domínio especializado se baseia em eventos que geralmente ocorrem nele e podem ser representados (Grygiel, 2017). Nessa mesma direção, Faber (2012) aponta para a vantagem de se utilizar a noção de *frames* para tornar explícito tanto o potencial semântico quanto sintático das unidades de conhecimento especializado que, necessariamente, incluem as descrições das relações conceptuais e o potencial combinatório de um termo.

Faber (2014) destaca que os *frames* aplicados à terminologia evidenciam as relações conceptuais hierárquicas e não-hierárquicas, com sua representação interna ou externa. Para a representação interna, os *frames* são utilizados nas relações lexicais e nos componentes encontrados nas definições terminológicas. Essas definições podem vir de informações de recursos de conhecimento especializado, assim como de um *corpus* compilado considerando o domínio específico a ser trabalhado. Na representação externa, busca-se representar as relações entre conceitos por meio de uma ontologia. Para isso, pode-se considerar modelos ou conceptualizações de algum domínio, compartilhados por um conjunto de usuários. Como afirma Faber (2014, p. 19),

uma ontologia de domínio específico, composta por conceitos e instâncias dentro de um determinado campo, juntamente com suas relações e propriedades, é um meio para armazenar e propagar o conhecimento especializado. Nesse sentido, a FBT propõe uma ontologia de base linguística, uma vez que seu desenho conceptual se baseia em informações extraídas de textos especializados e na estrutura de definições terminológicas.

A elaboração de uma ontologia nos moldes da *Frame-based Terminology* permite que se compreenda a terminologia, pensando nos diferentes contextos em que ela pode aparecer, assim como adiciona mais uma camada de informação linguística advinda das definições terminológicas. Esse ponto destacado por Faber (2014) vai ao encontro do que pretendemos neste trabalho, visto que a parte aplicada da tese é a construção de um *Knowledge Graph* (KG)<sup>19</sup> com *frames* semânticos. A etapa de arquitetar um KG com o modelo conceptual nos permitirá adicionar diferentes informações, como os termos a serem harmonizados nos diferentes contextos de uso e as definições terminológicas de cada unidade. A construção desse tipo de representação do conhecimento contribuirá não só para compreender os usos dos termos para a tarefa de harmonização de termos e conceitos como também para utilizá-la na automatização do trabalho terminológico, quando solicitado ao ChatGPT para desenvolvê-la. No que se refere ao uso do KG e dos *frames* no ChatGPT, sabe-se que, quando se insere a verbalização da representação do conhecimento no modelo de linguagem, ela potencializa o raciocínio do modelo, permitindo que ele forneça resultados com maior precisão e acurácia.

Um princípio importante discutido tanto na teoria fillmoreana quanto na Terminologia baseada em *frames* é a noção de contexto. Tendo em vista que as unidades lexicais servem como pontos de acesso para o conhecimento especializado (Faber, 2012), a visão enciclopédica que norteia a Semântica de *Frames* auxilia na compreensão da terminologia. Uma das formas de entender a estrutura do domínio está relacionada com o contexto no qual está inserida. Assim, Fillmore (1982, p. 119) afirma que “um ‘*frame*’, do modo como essa noção é empregada na descrição de significados linguísticos, é um sistema de categorias estruturado de acordo com um determinado contexto motivador”. Da mesma forma que Fillmore (1982, 1985) destaca a importância do contexto em sua teoria, a Terminologia baseada em *Frames* também faz essa relação quando diz que o significado depende do contexto (Faber, 2012). Assim, para Faber e Cabezas-García (2019, p. 199),

---

<sup>19</sup> Na seção 4.2, Representação do Conhecimento: Ontologias, *Knowledge Graphs* e *Frames*, abordaremos mais detalhadamente o que é um *Knowledge Graph* e a sua relação com a Ontologia.



Dado que os conceitos não podem existir no vácuo, eles são mais significativos quando relacionados entre si e integrados em configurações de conhecimento progressivamente mais complexas. A experiência baseada em *frames* envolve a aplicação de conhecimento armazenado derivado de contextos e situações semelhantes com o objetivo de entender eventos complexos e como lidar com eles.

Nesse sentido, as unidades lexicais dão acesso às estruturas não linguísticas, ou seja, ao conhecimento enciclopédico. Ainda sobre a importância do contexto para compreender a unidade especializada, Faber (2004) e Faber *et al.* (2005) mencionam que, em um domínio especializado, no qual os conceitos estão situados, a noção de *frame* é útil não apenas para a definição de conceitos, mas também para sua localização dentro do contexto no qual as categorias estão relacionadas entre si. À medida que o contexto se torna um ponto fundamental para analisar os termos do domínio especializado, facilita a compreensão dos usos e do comportamento da terminologia. Em nosso trabalho, esse é um ponto chave, visto que, para harmonizar termos variantes de diferentes domínios, é preciso considerar não apenas a definição e o termo *per se*, mas também o contexto e o local em que a unidade terminológica está inserida. No caso da harmonização de termos do ERP, se considerarmos todos os setores que compõem o *software*, como Recursos Humanos, Finanças, Compra e Vendas etc., um mesmo termo pode apresentar muitas variações a depender do setor/domínio em que ele está inserido. Para ilustrar a relevância de se considerar o contexto<sup>20</sup> na tarefa de harmonização, podemos pensar no setor de Recursos Humanos, que apresenta os termos ‘empregado’ e ‘operário’. As unidades, nesse caso, referem-se a alguém que trabalha para uma instituição/empresa. No entanto, se compararmos os termos ‘empregado’ e ‘operário’, podemos observar que a unidade ‘empregado’ é mais ampla e é utilizada em diversos contextos para denominar qualquer pessoa que trabalha em uma organização, independente da posição que ocupa na empresa. No entanto, isso não ocorre com o termo ‘operário’, que, geralmente, é utilizado para designar alguém que realiza um trabalho mais manual e braçal, em um contexto de chão de fábrica. Nesse sentido, considerar o contexto e compreendê-lo é crucial para a harmonização, uma vez que ele indicará todas as possibilidades de uso e aplicação da terminologia.

---

<sup>20</sup> Neste trabalho, entendemos “contexto” como tudo o que está relacionado aos termos, extrapolando as definições terminológicas e dicionarísticas. Consideram-se também as relações da terminologia com as diferentes áreas e setores do ERP, informações jurídicas, culturais, sociais, históricas advindas dos *corpora* compilados para esta pesquisa, relevantes para a seleção do termo principal, e das variantes na tarefa de harmonização de termos e conceitos.

Na Semântica de *Frames*, busca-se descrever uma cena abstrata ou uma situação, que, em muitos casos, tem como núcleo um verbo. Nessa perspectiva, o verbo desempenha um papel fundamental na organização e estruturação do *frame*, já que ele define uma ação, a qual indica um evento. Além disso, o verbo também estabelece as relações entre os elementos de *frames*. Como vimos no exemplo *transação\_comercial*<sup>21</sup>, o verbo auxilia na ativação do conhecimento contextual e contribui para a interpretação do significado, dependendo do contexto em que ele está inserido. Pelo viés da Terminologia, ainda que o uso de verbos seja importante para os estudos e as aplicações terminológicas, percebe-se que eles não recebem devida atenção, posto que os verbos estão mais relacionados à linguagem geral. Nas palavras de Faber e Cabezas-García (2019, p. 205),

na linguagem especializada, isto soa estranho porque os verbos raramente são considerados termos e, portanto, geralmente não são incluídos em recursos de conhecimento especializado (L’Homme, 1998; Buendía Castro, 2013). No entanto, os verbos da linguagem geral são cruciais para o significado porque geralmente são eles que relacionam conceitos em textos especializados.

Para Faber e Cabezas-García (2019), os verbos de linguagem geral ativam o conhecimento especializado. Em textos de especialidades, os verbos indicam como as unidades terminológicas são combinadas e utilizadas nos eventos do domínio. Dessa forma, as autoras afirmam que é imprescindível que se reflita mais sobre o papel semântico dos verbos em texto de domínio especializado, já que eles são as unidades que ligam a linguagem geral à linguagem especializada.

Outro ponto que vale salientar é a estreita relação entre a Terminologia baseada em *Frames* e os Estudos da Tradução. No tocante à aplicação do modelo de Fillmore (1982, 1985) para análise semântica em textos especializados, Faber (2012, p. 22) mostra que, “como a função geral dos textos de linguagem especializada é a transmissão do conhecimento, tais textos tendem a se adequar a modelos para facilitar a compreensão, assim como são caracterizados por uma repetição maior do que o habitual de termos, frases, sentenças e até mesmo parágrafos completos”. Nesse sentido, para os tradutores, quanto maior for a repetição de termos e estruturas, mais fácil se torna o trabalho tradutório. Isso ocorre, pois as ferramentas computacionais voltadas para a tradução, as *Computer-assisted Translation Tools* (CAT Tools) e as ferramentas

---

<sup>21</sup> O exemplo do frame *Transação\_Comercial* está descrito no capítulo 3, Linguagem e Cognição: A Semântica de *Frames*, na subseção 2.2 Semântica de *Frames*.

terminológicas memorizam as repetições e fornecem, posteriormente, a melhor solução para o tradutor. Logo, conectar um *plug-in* de gerenciamento terminológico em uma *CAT Tool* auxilia no fazer tradutório, dando suporte para tornar a tradução mais eficiente e rápida. Uma forma de aprimorar os fazeres tradutório e terminológico é desenvolver ferramentas baseadas em *frames* semânticos que contribuem para compreender o domínio e auxiliam na busca por dados da terminologia e da tradução.

No que tange à abordagem metodológica, conforme vimos na subseção *Semântica de Frames e suas aplicações*, a Terminologia baseada em *Frames* integra duas metodologias, a *bottom-up* e a *top-down*. A primeira corresponde ao uso de *corpora* multilíngues para extração de informações sobre o domínio a ser investigado, e a segunda diz respeito à inclusão de mais informações acerca do domínio retiradas de materiais referenciais e dicionários. Simultaneamente, o estudo do domínio indica as categorias mais genéricas que serão representadas em um evento. A partir disso, cria-se um modelo aplicável a todos os níveis de informação estruturada (Faber, 2012). Em sua metodologia, Faber (2012) mostra que a descrição dos domínios especializados é baseada em eventos, e cada um desses domínios fornece um *frame* para organizar os conceitos específicos. Dessa forma, há uma relação hierárquica e não-hierárquica na organização dos conceitos. Após essa etapa, elaboram-se modelos para inserir as relações conceptuais prototípicas, definições, dentre outras informações, para cada subdomínio do evento.

Além da discussão sobre a abordagem metodológica, na subseção *Semântica de Frames e suas aplicações* também apresentamos o recurso terminológico EcoLexicon, desenvolvido pelo grupo de pesquisa LexiCon, da Universidade de Granada, que segue os princípios da Terminologia baseada em *Frames*. Para além desse recurso online, o grupo LexiCon também tem desenvolvido outros estudos no domínio das ciências militares (Faber; León-Araúz, 2019, 2019b) e no domínio ambiental (León-Araúz; Faber; Pérez, 2008; León-Araúz; Reimerink; Faber, 2008; Reimerink; Faber, 2009; León-Araúz; Reimerink; Faber, 2009; Ureña; Faber; Buendía Castro, 2013; Ureña; Buendía Castro, 2017; Ureña; Moreno; Buendía Castro, 2017; Rojas-García; Faber, 2019; Gil-Berrozpe, 2020; Faber; León-Araúz, 2021).

Após a discussão acerca das teorias terminológicas, em especial da Terminologia baseada em *Frames*, trataremos, a seguir, da harmonização de conceitos e termos, apresentando tal abordagem metodológica e seu processo de aplicação.

### 3.3 Harmonização de conceitos e termos: definições e práticas

Dentro da Terminologia, seguindo a linha teórica wüsteriana, encontramos algumas abordagens metodológicas que visam atingir um certo nível de padronização dos termos. Dentre as metodologias descritas, a harmonização de termos e conceitos, concepção desenvolvida pela Organização Internacional para Padronização (*ISO – International Organization for Standardization*), refere-se à análise e à descrição de terminologias, objetivando recomendar usos de termos e conceitos para evitar inconsistências terminológicas e tornar a comunicação especializada mais eficiente. Tendo em vista que um dos objetivos deste trabalho é desenvolver uma metodologia semântico-terminológica que trate da harmonização de termos, inicialmente, discutiremos sobre a ISO/TC 37, comitê responsável pelo desenvolvimento de certificações no âmbito da Terminologia, traçando um breve histórico da organização e apontando para algumas questões acerca da padronização de termos. Em seguida, trataremos da variação terminológica, ponto chave na harmonização, visto que lidamos com termos variantes na tarefa. Posteriormente, compreenderemos melhor o que é a harmonização terminológica, apresentando essa abordagem, bem como descrevendo os passos para aplicar o processo de harmonização em um determinado domínio.

A busca pela padronização terminológica ocorreu, sobretudo, pelo processo de globalização e avanço científico e tecnológico. A globalização, o desenvolvimento mercantil e a criação de blocos econômicos fizeram com que a troca de informações em níveis comerciais, culturais e sociais aumentassem significativamente, impactando na produção e tradução de documentos técnicos (Krieger; Finatto, 2004). Para Cabré (1998), no século XIX, com o advento da tecnologia e o desenvolvimento de computadores, a busca por métodos de documentação e traduções técnicas fizeram com que surgissem os primeiros bancos de dados e, como consequência, iniciou-se a coordenação internacional de princípios de processamento terminológico, apresentando as primeiras abordagens para padronizar termos dentro de uma língua. Nesse sentido, a ISO/TC 37 acompanhou a nova virada tecnológica e passou a intensificar seus projetos que trabalham com gerenciamento terminológico e tradutório e suas tecnologias.

Como vimos na seção 3.1, Teorias Terminológicas: Da Teoria Geral da Terminologia à Terminologia Sociocognitiva, muitos pressupostos preconizados por Wüster (Wüster, 1968, 1974; Felber, 1979) foram fundamentais para a criação de organizações normalizadoras, como, por exemplo, a *International Organization for*

*Standardization* (ISO) (Waquil, 2017). As certificações da ISO, em especial os documentos relativos à Terminologia, são desenvolvidas pela ISO/TC 37, comitê técnico voltado para os aspectos terminológicos das línguas. A ISO foi criada em 1947, após a publicação da tese de doutoramento de Wüster, que tratou da padronização terminológica na área da engenharia (Felber, 1979). O trabalho desenvolvido pela Teoria Geral da Terminologia serviu como inspiração para o surgimento do comitê, conforme afirmam Krieger e Finatto (2004, p. 28),

O ideal de alcançar a padronização terminológica para garantir a perfeita intercomunicação científica e técnica em plano internacional, assim como o intuito de unificar os métodos de trabalho em Terminologia, está na origem da constituição de organizações internacionais voltadas a essas finalidades. Tais objetivos, preconizados por Wüster, motivaram até mesmo a criação do Comitê Técnico 37 da ISO (*International Standard Organization*), intitulado “Terminologia: princípios e coordenação”.

As ideias wüsterianas alicerçaram os princípios de normalização e de organização de muitas normas da ISO, como podemos ver no objetivo principal do comitê, que se ocupa da “padronização das descrições, recursos, tecnologias e serviços referentes à terminologia, à tradução, à interpretação e demais atividades relacionadas à linguagem no contexto multilíngue” (*International Organization for Standardization* (ISO), 2024)<sup>22</sup>. Além disso, segundo Waquil (2017), entre os anos de 1967 a 1973, por meio do Comitê Técnico 37, *Language and Terminology*, Wüster elaborou seis recomendações da ISO relativas à Terminologia: ISO 1087, *Terminology work and Terminology Science*; ISO 704, *Terminology work – Principles and methods*; ISO 860, *Terminology work – Harmonization of concepts and terms*; ISO 919, *Guide for the preparation of classified vocabularies (example of method)*; ISO 1149, *Layout of multilingual classified vocabularies*; e ISO 639, *Language code* (Waquil, 2017).

No que tange à organização do Comitê ISO/TC 37, encontramos cinco subcomitês: o ISO/TC 37/SC 1, que se volta para os princípios e métodos da Terminologia; o ISO/TC 37/SC 2, relativo aos aspectos de fluxo terminológico e codificação da língua; o ISO/TC 37/SC 3, que se ocupa dos recursos de gerenciamento de terminologias; o ISO/TC 37/SC 4, referente aos recursos de gerenciamento de língua; e o ISO/TC 37/SC 5, que trata das tecnologias ligadas à tradução e à interpretação.

---

<sup>22</sup> SCOPE. In: ISO/TC 37. [Genebra: International Standardization for Organization]. Disponível em: <https://www.iso.org/committee/48104.html> Acesso em: 9 jun. 2024.

Considerando que as recomendações da ISO, em especial as desenvolvidas por Wüster (1968, 1974), partem dos pressupostos da Teoria Geral da Terminologia (1979), cabe aqui discutirmos sobre a variação terminológica. A variação é um fenômeno comum na Terminologia, mas pouco tratado por Wüster e pela ISO, além de ser foco principal no nosso trabalho de harmonização. Pela perspectiva da Terminologia tradicional (Wüster, 1968, 1974; Felber, 1979), buscava-se eliminar as variações de termos e conceitos para se atingir uma comunicação unívoca. Tendo em vista a influência dos pressupostos de Wüster na criação das recomendações, a ISO 1087 (2019) estabeleceu que uma designação é uma expressão linguística ou não-linguística, que representa um conceito e que denota a um domínio. O termo é uma designação que representa um conceito geral por meio de uma expressão linguística. Como conceitua Biel (2023), quando um conceito é evocado por um termo, ele apresenta uma relação um-a-um (de forma e significado), e isso possibilita a univocidade ou monoreferencialidade, visto que só teríamos a relação entre um conceito e um termo. Fernández-Silva, Freixa e Cabré (2012) salientam que, na visão wüsteriana, se o objetivo era ter a relação termo-conceito no contexto especializado para se atingir a univocidade, buscou-se evitar a ambiguidade e administrar as variações lexicais, como a polissemia, homonímia e sinonímia.

Como também reitera Biel (2023), a língua é suscetível à variação, isto é, admite mudanças tanto diacronicamente como sincronicamente. Desde a metade dos anos 1990, a variação terminológica se tornou um fenômeno muito discutido na Terminologia, principalmente com o advento da tecnologia, que proporcionou as análises computacionais de termos, e com o crescimento de teorias terminológicas associadas às abordagens mais funcionais da linguagem (Biel, 2023). Segundo Fernández-Silva, Freixa e Faber (2012), a variação terminológica ainda é um campo pouco explorado, visto que, para o terminólogo, é um desafio descrever as variações e suas especificidades nas linguagens especializadas. Além disso, é complexo examinar as causas da variação e explorar os processos de criação e de transferência do conhecimento a partir das variantes.

Se, antigamente, o objeto de estudo da Terminologia era o conceito, como vimos na Teoria Geral da Terminologia (Wüster, 1968, 1974; Felber, 1979), posteriormente, passou-se a considerar o termo como o cerne da investigação terminológica. À vista disso, aceita-se a variação lexical e os ruídos que podem “afetar” a comunicação, como a sinonímia, polissemia, ambiguidade e a homonímia. Desse modo, observa-se que a

variação ocorre, sobretudo, pelas diferentes formas de nomear um conceito e por considerarem certas nuances conceptuais, sociais e pragmáticas na sua formação (Sánchez, 2011).

Fernández-Silva, Freixa e Faber (2012) afirmam que há dois tipos de variação: a variação denominativa, que afeta a forma, e a variação conceptual, que afeta o significado. Dada a relação muito próxima entre os dois tipos devido à ligação entre termo e conceito, a variação denominativa geralmente está na origem da mudança ou é resultado de alguma alteração conceptual. Logo, a variação denominativa é o tipo mais encontrado, uma vez que inclui diversos fenômenos como a sinonímia, a hiperonímia, a hiponímia ou os sinônimos próximos. Nessa variação, abrangem-se todas as diferenças formais, como um termo e o seu acrônimo ou dois termos com escritas diferentes. A título de exemplo, a sinonímia é um dos fenômenos mais discutidos, pois engloba todos os termos utilizados para representar o mesmo conceito (Fernández-Silva; Freixa; Faber, 2012).

Além da mudança lexical, a variação denominativa também apresenta diferenças semânticas (Fernández-Silva; Freixa; Faber, 2012). Do ponto de vista cognitivo, Sánchez (2011) parte da hipótese de que a multidimensionalidade do termo pode ser investigada como um fenômeno cognitivo dinâmico que delinea as formas das variantes terminológicas. Essa natureza dinâmica advém das bases do conhecimento especializado e da representação da forma contextualizada com que conceptualizamos o mundo. De acordo com Sánchez (2011, p. 183), sua proposta de variação terminológica considera três pontos fundamentais:

1. Nossas conceptualizações e verbalizações são situadas, muitas vezes moldadas pelos nossos públicos, relacionamentos e parceiros de comunicação;
2. O nosso pensamento e as nossas conceptualizações são moldados pelas características lógicas e externas dos objetos, mas também pela nossa formação cultural e pelo nosso conhecimento de mundo;
3. A forma como interagimos com o ambiente e seus objetos, propriedades, eventos e processos é dinâmica e configura a forma como nos referimos aos conceitos.

Nesse sentido, uma variante pode ser vista como uma forma semanticamente e conceptualmente associada ao termo principal (Sánchez, 2011; Daille, 2017). Vale destacar que um termo variante não é, obrigatoriamente, um sinônimo total do termo principal. A variante pode ser compreendida como uma unidade que é, de forma semântica e conceptual, similar ao termo principal, e os dois são ativados de acordo com

o contexto (Sánchez, 2011; Daille, 2017). O contexto, nessa perspectiva, é o ponto fundamental para acessar as variantes, visto que realça as sutis diferenças do significado a níveis semântico e pragmático. Sob o rótulo de variante, entendemos os termos variantes como um ‘tipo de’ ou ‘parte de’ um termo, como, por exemplo, ‘operário’ é um tipo de ‘empregado’.

Para o trabalho de harmonização de termos e conceitos, tratar da variação é crucial, visto que, na essência da tarefa, é preciso definir qual é o termo principal e quais serão as variantes. Como a nossa metodologia semântico-terminológica para a harmonização parte das definições da ISO 860:2007 sobre como harmonizar e adicionamos mais uma camada de informação cognitiva dos *frames*, torna-se primordial definirmos as variantes considerando aspectos terminológicos e cognitivos. Nesse sentido, neste trabalho compreendemos as variantes a partir da visão de Fernández-Silva, Freixa e Faber (2012), Sánchez (2011) e de Daille (2017). Os autores assumem o aspecto cognitivo da variação, especialmente quando evidenciam que as variantes podem designar o mesmo conceito, apresentando o mesmo papel semântico na definição, e podem ser consideradas como parte do mesmo evento, indicando uma relação entre elas.

Outro ponto fundamental na definição de variação se refere ao contexto. Esse aspecto também é de importância para esta tese, uma vez que estamos lidando com diferentes termos, em diferentes contextos, mas que designam o mesmo conceito. Para ilustrarmos a relevância de se considerar o contexto nessa tarefa, um dos conjuntos de termos que iremos harmonizar é composto por ‘empregado’, ‘funcionário’, ‘colaborador’, ‘membro do pessoal’, ‘trabalhador’, ‘trabalhador temporário’, ‘recurso externo’, ‘usuário de colaboração’ e ‘operário’. Tais termos são utilizados para se referir à pessoa que trabalha em uma empresa, instituição ou organização. No entanto, dependendo do contexto, seleciona-se um termo e não os outros. Como vimos anteriormente, se escolhermos usar o termo ‘operário’, por exemplo, precisamos ter em mente que ele está associado a alguém que desempenha um trabalho mais braçal no chão de fábrica, não incluindo todas as pessoas que designam outros tipos de funções. Como podemos observar, o contexto associado aos aspectos terminológicos e cognitivos, nesses casos, é vital para que possamos compreender e nos aprofundar na seleção do termo principal e das variantes.

Retomando as discussões acerca das recomendações da ISO, com relação à harmonização de termos e conceitos, essa tarefa se encaixa dentro da ISO/TC 37/SC 1,



através da certificação ISO 860:2007, e, no Brasil, pela certificação NBR 13790:1997. Como um dos nossos objetivos é apresentar uma metodologia semântico-terminológica que seja capaz de lidar com a harmonização terminológica de uma base de dados e automatizá-la com o ChatGPT, trataremos dos documentos desenvolvidos pelo subcomitê ISO/TC 37/SC 1, o qual elaborou a certificação referente à harmonização.

No que se refere aos documentos referenciais publicados pela ISO/TC 37<sup>23</sup> e suas subcomissões, destacamos a ISO 704:2009, que trata dos princípios e métodos do trabalho terminológico – no Brasil, ABNT NBR 13789:1997. Além dessa, outra recomendação importante é a ISO 1087/1:2000, que se refere à teoria e às aplicações de vocabulário. A ISO 12618:1994 volta-se para o design, a implementação e o uso de sistemas de gerenciamento terminológico em aplicativos de computador. A ISO/AWI 12616, documento ainda em desenvolvimento, lidará com o trabalho terminológico de apoio à comunicação multilíngue (fundamentos para a orientação tradutória e terminográfica). Por fim, evidenciamos a ISO 860:2007 – no Brasil, NBR 13790:1997 – , atinente ao trabalho terminológico de harmonização de termos e conceitos. Vale destacar que a ISO 860:2007 é a segunda edição da recomendação acerca da harmonização de termos e conceitos. Ela está disponível nas línguas inglesa, francesa e russa. Com relação à tradução para o português, a ABNT disponibilizou apenas o documento NBR 13790:1997, isto é, não há, em língua portuguesa, a segunda edição de 2007 do documento.

Antes de nos aprofundarmos nas questões acerca da harmonização terminológica, é importante diferenciar algumas atividades que possam estar associadas à harmonização, como a planificação e a normalização. A concepção de planificação linguística pode ser compreendida quando países bilíngues ou multilíngues, como, por exemplo, Canadá, Espanha, Noruega e Islândia, passaram a adotar políticas linguísticas com a intenção de cultivar a língua materna e mantê-la “pura”. Nesse sentido, Faulstich (1998, p. 248) afirma que

o conceito de planificação linguística se apoia em um projeto linguístico coletivo. [...] Normalmente, a planificação decorre de um esforço conjunto para o estabelecimento de uma política linguística nacional. Nesse sentido, nós podemos dizer que a planificação é regulamentada pelas disposições jurídicas que, em matéria de língua, acabam por se constituir em um conjunto de regras legisladoras.

---

<sup>23</sup> SCOPE. In: ISO/TC 37. [Genebra: *International Standardization for Organization*]. Disponível em: <https://www.iso.org/committee/48104.html> Acesso em: 9 jun. 2024.

A concepção de planificação está relacionada à criação de políticas linguísticas, em especial aquelas voltadas para a valorização e preservação da língua, seja por motivação política, cultural, social, seja pela extinção da língua. Atualmente, a planificação linguística segue alguns princípios básicos, como manter o purismo linguístico. Contudo, para ser possível compartilhar o conhecimento especializado, os países que têm políticas linguísticas voltadas para a planificação desenvolvem recursos terminológicos com o intuito de divulgar determinado domínio e, ao mesmo tempo, manter viva sua língua (Picht, 2007).

Faulstich (1998) aponta para dois tipos de normalização: a técnica e a linguística. A primeira, a normalização técnica, está em consonância com os preceitos definidos pela ISO, sendo seu objetivo contribuir para a comunicação eficaz, o progresso mercantil e industrial, e facilitar a negociação entre países, organizações e empresas. Além disso, a normalização técnica internacional é fundamental para “aumentar a confiança do consumidor e usuário em produtos e serviços que estejam de acordo com seus requisitos por empresas que satisfaçam aos padrões de Gestão da Qualidade adotados pelo mercado” (Faulstich, 1998, p. 258). O segundo tipo de normalização, a linguística, considera a gramática da língua e os usos que fazem dela, refletindo sobre o funcionamento da língua no que tange à comunicação oral e escrita. Essa prática busca “aparelhar as línguas para todas as formas de expressão” (Krieger; Finatto, 2004, p. 29).

A harmonização terminológica é uma prática que se propõe a melhorar a comunicação entre países, organizações e instituições. Para Pavel e Nolet (2001), o processo de harmonizar termos, geralmente, é desenvolvido por uma empresa, órgão ou outra entidade que deseja adequar usos terminológicos em determinado domínio. A ISO divide a harmonização de termos e conceitos em duas partes. A primeira se refere à harmonização de conceitos, e a segunda à harmonização de termos. A *International Organization for Standardization* (2007) conceitua a etapa de harmonização de conceitos como uma tarefa que lida com o estabelecimento de um correspondente entre dois ou mais conceitos relacionados ou sobrepostos, que apresentam diferenças profissionais, técnicas, científicas, sociais, econômicas, linguísticas, culturais, dentre outras, para que se eliminem ou se reduzam pequenas divergências entre eles. Nesse sentido, busca-se descrever um conceito harmonizado a partir da definição intencional, a qual indica a posição do conceito no sistema harmonizado de conceitos. A harmonização de termos, por sua vez, está ligada à atividade de selecionar designações

para um conceito harmonizado, em um contexto multilíngue ou dentro de uma mesma língua (*International Organization for Standardization, 2007*). Caso a harmonização ocorra dentro de uma mesma língua, os termos harmonizados serão considerados como sinônimos ou variantes. Se a harmonização for feita entre línguas diferentes, é preciso pensar na equivalência entre os termos (ABNT, 1997). Logo, equivalência, nos termos da ISO 860:2007, é a relação entre os termos em diferentes línguas que apresentam o mesmo conceito.

Apesar de ser um procedimento bem definido, muito se questiona até que ponto a harmonização terminológica pode determinar e consagrar o uso de termos. Nesse sentido, Barros (2004, p. 88) afirma que

[...] a harmonização terminológica procura adequar, modificar ou consagrar o uso de determinados termos a situações ou contextos. Pode ser processada no âmbito de uma empresa ou outro tipo de órgão e constitui um meio caminho entre a normalização (mais impositiva, feita por especialistas da área, normalmente por organismos nacionais ou internacionais) e a recomendação pura e simples.

Com isso, observa-se que a harmonização de termos não é caracterizada pela imposição do uso de determinado termo sobre outro, mas pela sugestão de certos usos terminológicos em contextos distintos. Considerando que essa atividade não é impositiva, como a padronização e a normalização, Pavel e Nolet (2001) listam algumas atividades que fazem parte da harmonização terminológica, como, por exemplo, remover termos duplicados ou incorretos, confirmar o uso terminológico, propor novos termos, desaconselhar o uso de falso sinônimos ou variantes que geram confusão, recomendar o uso de certos termos e lidar com termos contraditórios. Como as autoras tratam da harmonização terminológica no contexto empresarial e político, tais ações devem ser conferidas, validadas e aprovadas por membros de um comitê ou gestores da empresa. A partir do desenvolvimento dessa prática, é possível ter maior precisão conceptual, realizar correções linguísticas, adequar a terminologia de acordo com a situação comunicativa e tornar a comunicação mais eficiente (Pavel; Nolet, 2001).

A seguir, apresentaremos a harmonização de conceitos e termos desenvolvida pela ISO/TC 37, definindo tal abordagem e apresentando os processos metodológicos elaborados pelo comitê da empresa.

### 3.3.1 Harmonização de conceitos e termos: metodologia e aplicação

Considerando a interdisciplinaridade da Terminologia e a constante evolução terminológica nas mais diversas áreas, percebe-se que conceitos e termos também se modificam entre as línguas e nas comunidades linguísticas, especialmente pelos aspectos históricos, geográficos, sociais e econômicos (ABNT, 1997). Como resultado, é necessário reestruturar o sistema de conceitos e termos e buscar harmonizá-lo, evitando, assim, prejuízos comunicacionais entre as comunidades tanto no nível nacional como no internacional (*International Organization for Standardization*, 2007). Como vimos na seção anterior, a harmonização de conceitos e termos não é uma tarefa impositiva, mas serve para adequar e sugerir o uso de termos e conceitos em um determinado domínio e em dado contexto. Contudo, antes de passarmos para a descrição do processo de harmonização terminológica, é importante discutir a natureza do conceito, noção importante para essa abordagem.

Conforme descrito na ISO 704:2022, nem todos os objetos no mundo apresentam uma denominação e, por meio da observação e da conceptualização, podemos categorizar os objetos e transformá-los em conceitos. Os conceitos são construções mentais ou unidades de pensamento que serão representados de diferentes formas na comunicação. Nesse sentido, a construção do conceito segue a ordem de objeto → conceito → comunicação. Para a *International Organization for Standardization* (2022), o conceito não deve ser identificado como um objeto abstrato ou imaginado, visto que um objeto dessa natureza é observado e conceituado mentalmente para, então, ter uma designação atribuída ao conceito e não ao objeto em si. Contudo, para a norma elaborada pela ISO, “a ligação entre um objeto e sua designação ou definição é feita através do conceito, um nível mais alto de abstração” (*International Organization for Standardization*, 2022, p. 2). É importante salientar que o tratamento do conceito, para a ISO, é feito a partir da linguagem específica e em um contexto especializado.

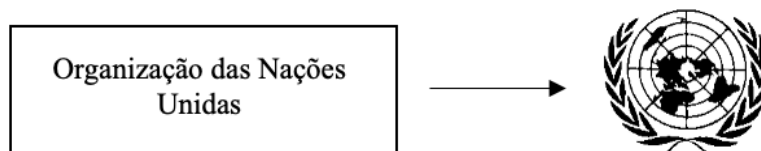
Tendo em mente que estamos tratando de conceitos em contexto e na linguagem específica, o conceito, nesse sentido, deve ser compreendido como uma unidade de pensamento, bem como uma unidade de conhecimento. Por essa perspectiva, quando o conceito está relacionado a um único objeto, chamamos de conceito individual, e este é representado como uma designação<sup>24</sup> ou um símbolo, por exemplo, Internet, *World*

---

<sup>24</sup> No documento ISO 704:2000, usa-se *appellation*. Para este trabalho, utilizamos a tradução de *appellation* como designação.

*Wide Web* e Estátua da Liberdade (ISO, 2000). No exemplo da Figura 19, podemos observar a designação Organização das Nações Unidas e seu símbolo.

Figura 19 - Relação entre designação e símbolo de Organização das Nações Unidas

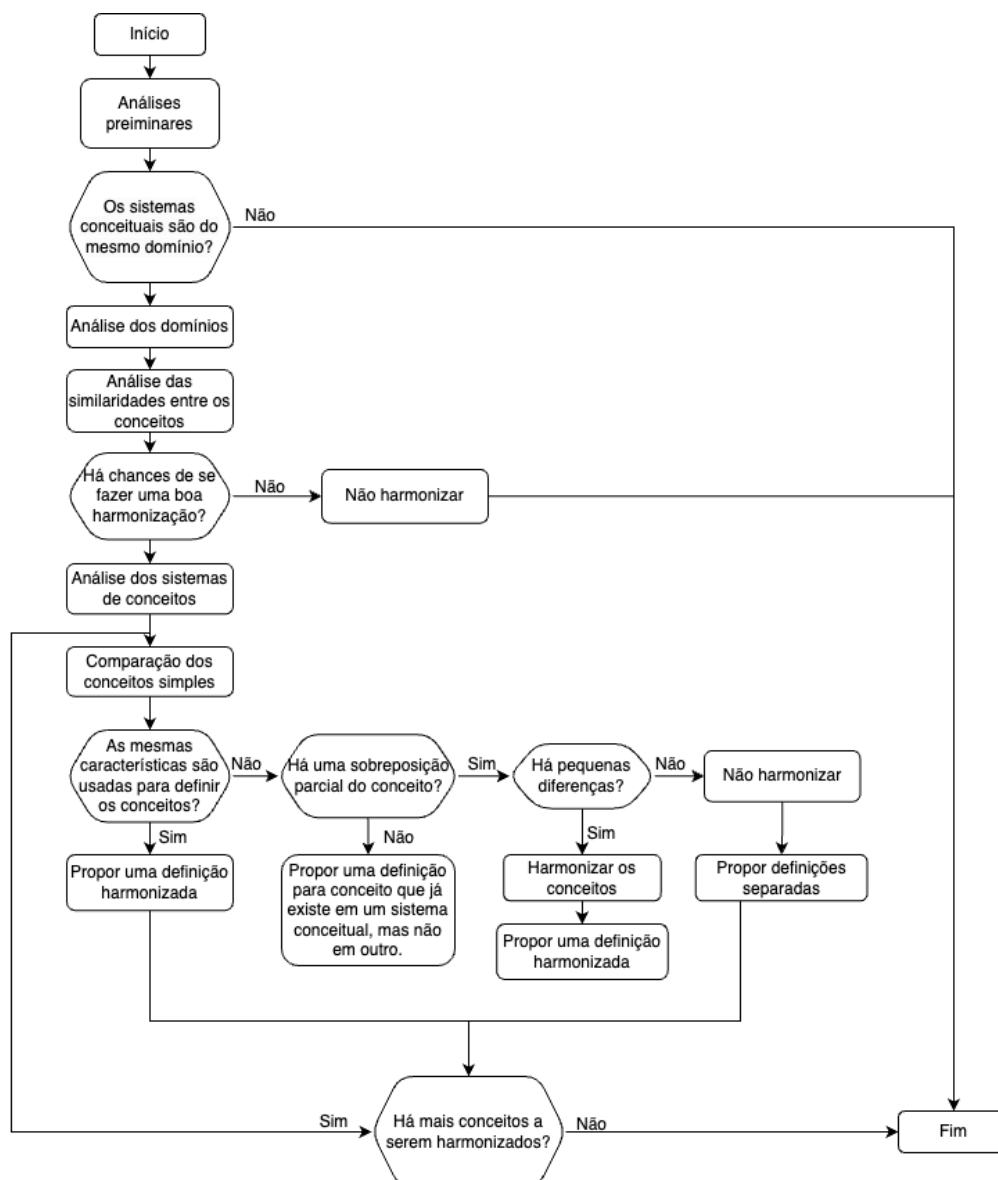


Fonte: Elaborado pela autora.

Caso o conceito represente dois ou mais objetos, chamamos de conceito geral, e sua designação assume a forma de um termo, por exemplo, “capital”, na área da Economia, “liquidez”, na área de Investimentos, e “navegador” (*browser*), na área da Computação (*International Organization for Standardization*, 2019, 2022). No contexto especializado, o conceito pode assumir diversas formas. Tal como na linguagem natural, o conceito pode ser visto como termo, designação, definição, dentre outros; na linguagem artificial, o conceito assume a forma de código ou fórmula, ícones, figuras, diagramas, dentre outras possibilidades. Nesse sentido, o conceito é fundamental para organizar o conhecimento, fornecendo recursos para identificar e reconhecer o objeto e agrupá-lo em unidades significativas na terminologia (*International Organization for Standardization*, 2022).

Conforme está descrito nos documentos ISO 860:2007 e NBR 13790:1997, a harmonização de termos e conceitos segue uma abordagem metodológica comparativa, a fim de verificar se realmente é possível aplicar a harmonização terminológica em um determinado domínio. É importante destacar que o processo de harmonização de termos e conceitos acontece em duas etapas: a primeira trata da harmonização de conceitos para, depois, ser possível harmonizar os termos (segunda etapa). Assim, o primeiro passo para a harmonização condiz com a comparação entre os sistemas de conceitos em uma ou mais línguas da área a ser trabalhada. Entretanto, não é permitido que se induza uma transferência direta entre conceitos que não são similares. Nesse sentido, a certificação proporciona um caminho que permite ao especialista refletir se é possível harmonizar os conceitos, conforme ilustra a Figura 20.

Figura 20 - Procedimentos para a harmonização de termos e conceitos



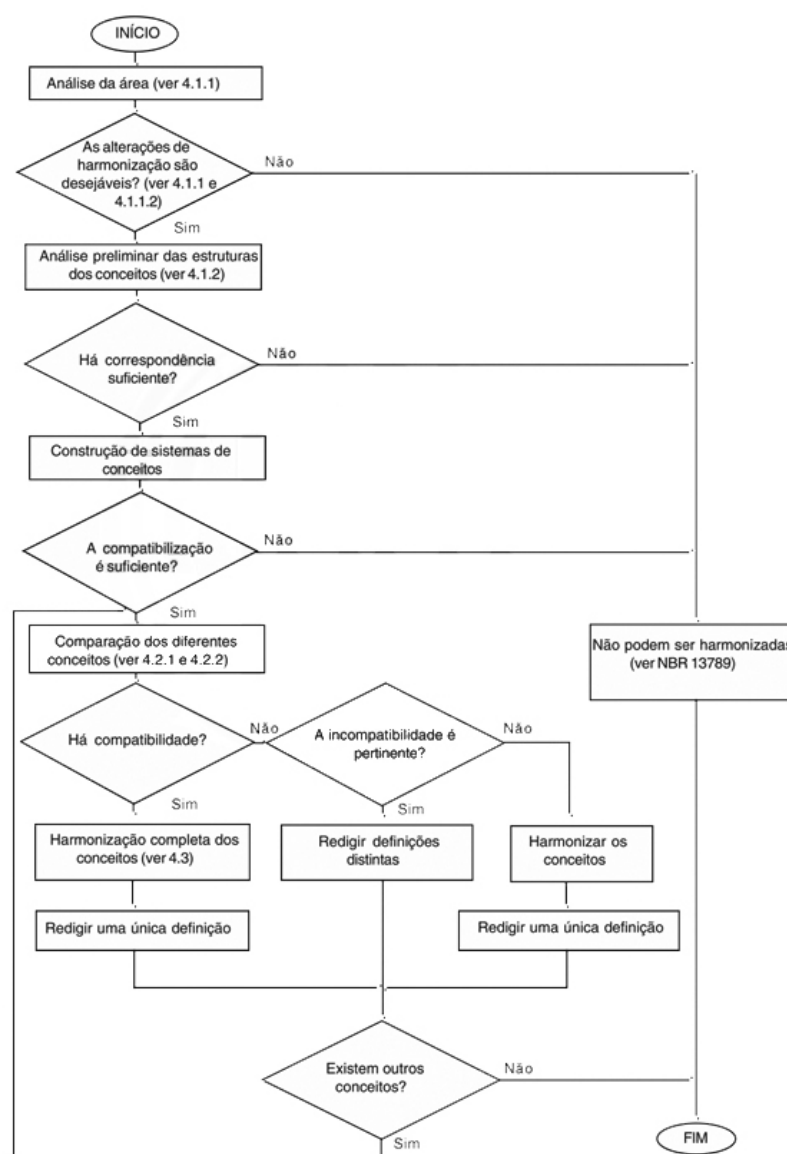
Fonte: Adaptado de *International Organization for Standardization* (2007).

Na Figura 20, podemos observar que a ISO apresenta diversos passos para que a harmonização seja feita da melhor forma possível. De início, analisam-se preliminarmente os dados a serem harmonizados. Em segundo lugar, observa-se se os sistemas de conceitos pertencem ao mesmo domínio e, caso não pertençam, a harmonização não deverá ser realizada. Se os conceitos forem do mesmo domínio, faz-se uma análise dos domínios e se analisam as similaridades entre os conceitos para, depois, observar se há chances de aplicar a harmonização. Caso a análise não apresente um bom resultado, não se deve harmonizar os conceitos. Caso o resultado seja positivo, deve-se, então, analisar os sistemas de conceitos e comparar os conceitos simples. Posteriormente, deve-se investigar se as mesmas características são utilizadas para

definir os conceitos, se há uma sobreposição parcial entre os conceitos e se há pequenas diferenças. Considerando as possibilidades, deve-se ou não harmonizar os conceitos,

Como dito anteriormente, a ABNT (1997) não traduziu a recomendação ISO 860:2007. Desse modo, podemos perceber algumas diferenças entre os procedimentos de harmonização do documento de 1997 para o de 2007. Para fins de comparação, na Figura abaixo, podemos observar o procedimento de harmonização apresentado pela ABNT (1997).

Figura 21 - Procedimento de harmonização NBR 13790:1997



Fonte: ABNT (1997, p.3).

Analisando os dois procedimentos (Figuras 20 e 21), podemos observar que a ISO (2007) adicionou etapas ao processo de harmonização. O diagrama da ABNT

(1997) inicia com a avaliação da área a ser harmonizada para, depois, questionar se as alterações de harmonização são desejáveis e, por fim, examinar preliminarmente as estruturas dos conceitos. No documento da ISO 860:2007, parte-se das análises preliminares acerca da harmonização. Em seguida, reflete-se se os sistemas de conceitos são da mesma área para, então, explorar o domínio e, por fim, verificar as similaridades entre os conceitos. Nessa primeira parte, podemos ver que as mudanças feitas na ISO 860:2007 buscam trabalhar de forma mais aprofundada na harmonização dos conceitos, realizando uma análise minuciosa do domínio e da estruturação do sistema conceitual.

Além dessas adições, podemos observar que a NBR 13790:1997 apresenta uma etapa que não é tratada na ISO 860:2007. Como descrito pela ABNT (1997), após a análise dos conceitos, investiga-se se há correspondência suficiente para realizar a harmonização. Depois, parte-se para a elaboração do sistema de conceitos e, em seguida, reflete-se novamente se a compatibilidade entre os sistemas é suficiente para harmonizar os conceitos. Essa etapa na ISO 860:2007 não é tratada da mesma forma. Como podemos ver, após o estudo da similaridade entre os conceitos, verifica-se se as chances de aplicar a harmonização são boas. Caso não seja possível, não será aplicada a harmonização.

A última parte da harmonização apresenta divergências entre as recomendações. Enquanto a NBR 13790:1997 discute se, mesmo depois de todas as análises, há compatibilidade para harmonizar os conceitos, então, deve-se realizar a harmonização completa dos conceitos e redigir uma definição. Caso não haja compatibilidade, questiona-se se a incompatibilidade é pertinente. Se ela é pertinente, recomenda-se redigir definições distintas. Caso não seja pertinente, sugere-se harmonizar os conceitos e redigir uma única definição. A ISO 860:2007 adiciona novos pontos a serem refletidos nesse processo. Após a comparação entre os conceitos, averigua-se se as mesmas características são utilizadas para definir os conceitos. Caso elas sejam as mesmas, recomenda-se propor uma definição harmonizada. Se não apresentar as mesmas características, o especialista deve se perguntar se há uma sobreposição parcial do conceito. Nesse caso, se não houver uma sobreposição, sugere-se propor uma definição para o conceito que já existe em um sistema, mas não em outro. Se houver uma sobreposição, examina-se se há pequenas diferenças entre eles. Caso haja, recomenda-se harmonizar os conceitos e propor uma definição harmonizada. Caso não haja, sugere-se que não se harmonizem os conceitos, recomendando, então, a proposição de definições separadas.



A partir dessa comparação entre as recomendações, podemos observar que o procedimento da ISO 860:2007 explora mais as análises conceituais e traz um maior detalhamento acerca de como lidar com as características do conceito. Para esta tese, consideraremos o procedimento descrito pela ISO 860:2007, visto que ela é a versão atual e apresenta diversos passos analíticos para a harmonização, tornando o processo mais completo e eficiente.

Aprofundando-nos na descrição do procedimento de harmonização (ISO 860:2007), observamos, na Figura 20, que há um longo processo para atingir a harmonização de um domínio específico. A partir de um estudo de viabilidade, a ISO propõe que o especialista investigue as semelhanças e diferenças entre os conceitos e o sistema de conceitos, a fim de atentar para a possibilidade de aplicar a harmonização (*International Organization for Standardization, 2007*). No que diz respeito às semelhanças e às diferenças entre conceitos, o órgão estabelece que conceitos dentro de um assunto em uma língua ou entre duas comunidades linguísticas são mais semelhantes e, conseqüentemente, a harmonização é mais fácil de ser feita. No Quadro 1, podemos observar as condições de semelhança e diferença dentro do domínio para garantir a harmonização.

Quadro 1 - Requisitos de semelhança e diferença para harmonização de termos e conceitos

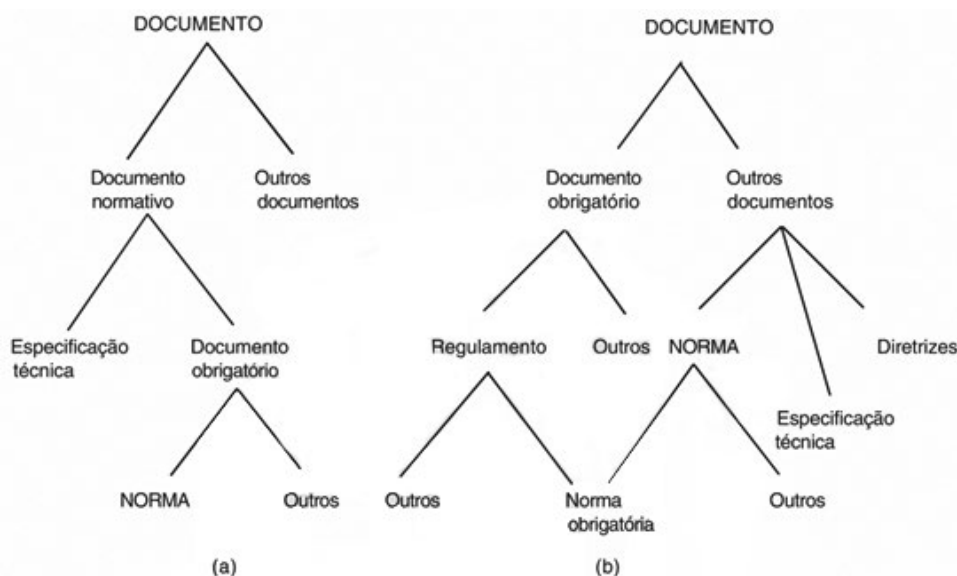
Semelhanças	Diferenças
(i) Quais características os conceitos de cada sistema de conceitos têm em comum (entre diferentes línguas ou na mesma língua); (ii) Quais características dos conceitos se diferem de um sistema de conceito para outro (entre diferentes línguas ou na mesma língua); (iii) Quais características são essenciais para cada conceito-chave.	(i) O número de conceitos incluídos no processo de harmonização; (ii) As relações entre os conceitos; (iii) A profundidade da estrutura de conceitos; (iv) Os critérios de subdivisão utilizados no desenvolvimento do sistema conceitual.

Fonte: Adaptado de *International Organization for Standardization (2007)*.

Após comparar as semelhanças e as diferenças entre os conceitos dentro do sistema (Quadro 1), é possível ver o potencial do domínio para aplicar a harmonização de termos e conceitos. O próximo passo descrito pelo documento condiz com a análise preliminar do sistema de conceitos, isto é, se as semelhanças entre os conceitos são significativas a ponto de poder, então, dar início ao processo de harmonização.

A partir dessa comparação, os sistemas conceituais devem ser montados em separado, objetivando relacionar os mesmos tipos de características entre os sistemas. Somente após essa etapa será possível observar se os sistemas são adaptáveis para aplicar a harmonização. Para exemplificar essa etapa, o documento NBR 13790 apresenta um exemplo de comparação entre sistemas, conforme a Figura 22.

Figura 22 - Dois sistemas de conceitos diferentes



Fonte: ABNT (1997, p. 4).

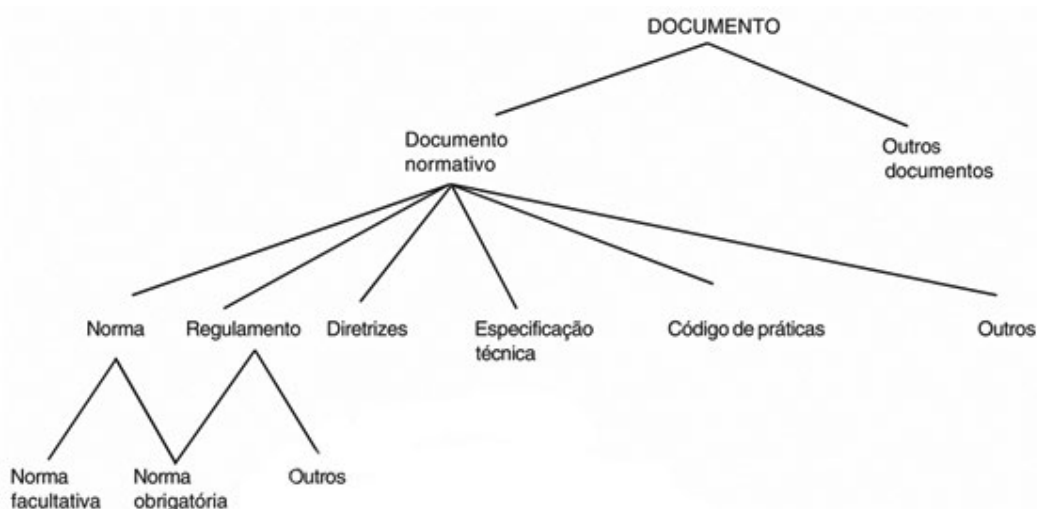
Na Figura 22, podemos observar que as estruturas não são compatíveis, visto que o sistema (b) apresenta ramificações na sua estrutura não descritas no sistema (a). Conseqüentemente, não há uma correspondência completa entre os conceitos dos sistemas (a) e (b). Os pontos compatíveis, comuns aos dois sistemas, são passíveis de serem harmonizados; os pontos sem correspondência em um dos sistemas não poderão ser harmonizados.

No que tange à harmonização de conceitos únicos (*single concepts*), deve-se realizar uma análise comparativa das suas características e não das suas designações. Nesse sentido, comparam-se apenas os atributos de cada conceito, desconsiderando as definições e os termos. Quando os conceitos são iguais, com características idênticas, em diferentes sistemas conceituais, a *International Organization for Standardization* (2007) indica que não há a necessidade de se harmonizarem os conceitos. Porém, com relação à harmonização de termos, talvez seja necessário aplicar a harmonização, atentando para a designação de cada conceito, na tentativa de estabelecer os sinônimos e variantes.

Outra possibilidade é lidar com a sobreposição parcial de conceitos. Nesse ponto, é preciso considerar se algumas das características fundamentais dos conceitos são diferentes, indicando divergências nos níveis intencional e extensional de cada conceito. Para essa análise, os especialistas devem determinar se as diferenças são pequenas ou consideráveis. Se as diferenças são pequenas, apenas um conceito pode ser utilizado. A *International Organization for Standardization* (2007) define as pequenas diferenças como resultados de alterações nos âmbitos profissional, técnico, científico, social, econômico, linguístico ou cultural. No caso de as diferenças serem consideráveis, será preciso lidar com dois ou mais conceitos. Desse modo, as características de cada conceito devem ser definidas, e diferentes posições devem ser atribuídas a cada conceito no sistema conceitual.

Com relação à determinação das diferenças e à continuidade da harmonização, caso as diferenças sejam pequenas, isto é, não haja uma variação que afete o conceito, o processo de harmonização estará concluído. Caso as diferenças sejam consideráveis, indicando mudanças de intensão ou extensão do conceito, então, é necessário atentar para outras condições a serem aplicadas. No caso de as diferenças observadas serem relevantes, apontando para a existência de vários conceitos diferentes e cada conceito deve ser definido para ocupar seu lugar no sistema harmonizado. A fim de ilustrarmos a harmonização de conceitos, podemos verificar, na Figura 23, como se estrutura um sistema já harmonizado.

Figura 23 - Exemplo de sistema harmonizado



Fonte: ABNT (1997, p. 5).

Na Figura 23, o sistema harmonizado mostra as relações entre os conceitos e como a harmonização foi aplicada, especialmente, se compararmos as Figuras 22 e 23. A Figura 23 se refere às diferenças entre as definições serem significativas. Neste caso, somente um conceito é o suficiente e ele deve ser definido de forma coerente no sistema. Vale salientar que, de acordo com a certificação, nem todos os sistemas de conceitos exclusivos apresentam um conceito correspondente na harmonização e, portanto, é justificável a opção de preencher essa lacuna devido à falta de correspondente. Por fim, para efetivar a harmonização de conceitos, é necessário que haja um consenso entre todos os especialistas quanto às características essenciais dos conceitos e se elas devem fazer parte da harmonização.

Com a primeira parte da harmonização pronta (harmonização de conceitos), partimos para a harmonização de termos, visto que essa etapa tem uma relação próxima com o trabalho realizado com os conceitos. De acordo com a ISO (2007, p. 11), “a harmonização de termos é direcionada para questões de seleção de termos, estabelecimento de equivalências, sinonímia e variação de termos. A harmonização dos termos só é possível se os conceitos subjacentes forem praticamente idênticos”. Considerando isso, percebe-se que a etapa de harmonizar a terminologia conta com três etapas: (i) seleção de termos; (ii) definição de equivalentes, sinônimos e variantes; e (iii) definição de variação entre termos em uma mesma língua.

Com relação à seleção de termos, a *International Organization for Standardization* (2007, p. 11) parte do princípio de que, “se os termos forem criados em uma língua individual, a estrutura do sistema de conceitos deverá ser refletida na formação dos termos sempre que possível”. Essa etapa pode ser aplicada, por exemplo, em domínios, sobretudo das Ciências Biológicas, que apresentam uma longa tradição na formação de termos, como na Química, na Biologia e na Medicina. Além disso, essa etapa de seleção também discute sobre a aplicação da tarefa em campos em que a cooperação internacional é comum e evidencia uma terminologia bem estabelecida. Nesses casos, os termos devem refletir as mesmas características do conceito (ISO, 2007). Ainda nessa parte, a ISO (2007) indica algumas possibilidades de substituição de termos que o especialista pode encontrar durante a etapa de seleção, como ilustra o Quadro 2.

Quadro 2 - Exemplos de substituição de termos definidos

	Substituição de termos estabelecidos	Termos formados pela tradução dos elementos do termo de um termo estrangeiro	Termos formados utilizando formas idênticas ou similares entre diferentes línguas
Definição	Algumas vezes, a harmonização pode sugerir que se substitua uma terminologia bem definida. Neste caso, deve-se considerar as vantagens e desvantagens de harmonizar os termos.	Para a nomeação de novos termos, deve-se utilizar as características idênticas do conceito a fim de se atingir uma correspondência terminológica internacional. Neste caso, o termo pode ser formado por meio da tradução do termo ou dos aspectos morfológicos do termo estrangeiro.	Termos harmonizados podem ser formados a partir de formas idênticas ou similares de termos equivalentes ou de elementos de termos entre diferentes línguas. Nesse caso, as formas podem ser consideradas idênticas ainda que apresente alguma variação ortográfica, utilize distintos padrões de palavras, a pronúncia seja diferente ou o alfabeto utilizado não seja o mesmo.
Exemplo	Termo: Manipulação genética Termo substituto: Engenharia genética	Inglês: <i>Acute bacterial myocarditis</i> . Português: Miocardite bacteriana aguda.	Inglês: <i>Philosophie</i> Português: Filosofia.

Fonte: Adaptado de *International Organization for Standardization* (2007, p. 12).

No Quadro 2, podemos observar que a ISO apresenta três casos distintos de substituição de termos em que a seleção de termos deve ser ponderada. No primeiro caso, Substituição de termos estabelecidos, a aplicação da tarefa pode conduzir ao trabalho com termos bem definidos. Contudo, consideram-se vantagens desvantagens de harmonizar tais termos, visto que podem gerar problemas de comunicação, dependendo da definição do termo na comunidade científica. A título de exemplo dessa harmonização, sugere-se pensar se há conotações negativas no termo original para que se substitua por um termo mais aceitável. No segundo caso, Termos formados pela tradução dos elementos do termo de um termo estrangeiro, considerando o contexto internacional, sugere-se que a nomeação de novos termos seja feita quando as características do conceito são idênticas para que se atinja uma correspondência do termo a nível internacional. Para o último caso, Termos formados utilizando formas idênticas ou similares entre diferentes línguas, recomenda-se que os termos harmonizados possam usar formas idênticas ou similares de termos equivalentes ou de elementos de termos entre diferentes línguas. Nessa situação, ainda que se encontrem variações ortográficas de padrões de palavras, de pronúncia ou de diferentes alfabetos, identifica-se que os termos são similares ou idênticos.

A segunda etapa da harmonização se refere à definição de equivalentes, sinônimos ou termos variantes. Segundo a ISO (2007), essa parte busca compreender e analisar os termos entre línguas ou na mesma língua. Quando a harmonização é aplicada entre diferentes línguas, estabelecem-se os equivalentes da terminologia. No caso de a harmonização ser feita na mesma língua, sugere-se definir os sinônimos ou termos variantes da terminologia. Nesse ponto, deve-se atentar para os falsos cognatos devido às semelhanças gramaticais superficiais (ISO, 2007).

A equivalência ocorre entre termos que designam o mesmo conceito nas diferentes línguas, como, por exemplo, o equivalente de ‘book’ é ‘livro’. Quando se trabalha com apenas uma língua, os sinônimos ou quase-sinônimos (*quasi-synonymy*) ocorrem entre termos que designam o mesmo conceito. Enquanto na sinonímia o contexto de uso é intercambiável, na quase-sinonímia, pode-se utilizar o termo em alguns contextos e em outros não (*International Organization for Standardization*, 2007). Para ilustrar a sinonímia, podemos pensar em ‘casamento’ e ‘matrimônio’, visto que os termos se referem à mesma situação de união entre duas pessoas e podem ser utilizadas de forma intercambiável. No caso de quase-sinonímia, podemos considerar os verbos ‘discutir’ e ‘brigar’. O verbo ‘discutir’ está mais ligado ao debate ou à troca de ideias e argumentos, enquanto ‘brigar’, que apresenta uma conotação mais negativa, refere-se à luta física ou a uma discussão mais acalorada, em que, por exemplo, pode haver troca de ofensas.

Quando tratamos da definição de termos variantes em uma mesma língua, deparamo-nos com a variação entre termos que designam o mesmo conceito na língua. Neste caso, a *International Organization for Standardization* (2007) apresenta as variantes geográfica, ortográfica, de registro e de estilo. A fim de ilustrarmos as diferenças entre os quatro tipos de variação, podemos pensar no uso de ‘mandioca’ e ‘aipim’ como variante geográfica; ‘assobiar’ e ‘assoviar’ para a variante ortográfica; a formalidade e informalidade com relação à variação de registro em ‘Prezado Senhor, gostaria de solicitar sua presença na reunião de amanhã’ e ‘Ei, pode vir à reunião amanhã?’; e, por fim, aspectos estilísticos quando descrevemos uma situação, como em ‘As flores eram de um vermelho intenso e exalavam um perfume doce’ e ‘As flores eram vermelhas e perfumadas’.

Após as etapas de seleção de termos, definição de equivalentes, sinônimos e variantes e a definição de variações entre termos na mesma língua, a *International Standard Organization* (ISO) aborda, rapidamente, o planejamento terminológico.

Nesse ponto, a ISO alerta para as possíveis diferenças entre o planejamento terminológico em uma determinada comunidade linguística e a comunidade internacional, uma vez que elas podem utilizar termos diferentes e, nesse caso, a harmonização não será aplicada para todos. Pensando nisso, a *International Organization for Standardization* (2007, p. 14) destaca que utilizar um termo supranacional “[...] facilita a comunicação entre comunidades linguísticas, mas, por outro lado, também pode ser um obstáculo à comunicação dentro de uma comunidade linguística”. Desse modo, a ISO ressalta que a decisão acerca da forma do termo deve ser determinada pelas comunidades linguísticas individuais, considerando a sua necessidade de harmonização.

O trabalho de padronização e harmonização da terminologia, apesar de sofrer duras críticas por seguir a linha wüsteriana, também apresenta alguns pontos positivos, sobretudo na área tecnológica atrelada aos estudos terminológicos e tradutórios (*machine Translation*). Para L’Homme (1994), é interessante aplicar uma padronização e harmonização em *softwares* de gerenciamento de termos, pois os tradutores recorrem a dicionários para buscar equivalentes de tradução e, então, selecionar qual é a melhor opção para se utilizar em determinado contexto. Além disso, vale salientar que, além da participação de especialistas e terminólogos, o envolvimento de tradutores durante o processo de harmonização e padronização é fundamental, uma vez que “os tradutores são qualificados de forma única para identificar contextos, estabelecer prioridades de frequência e apontar problemas que afetam os resultados da máquina” (L’Homme, 1994, p. 125).

Tendo em mente as informações discutidas sobre a ISO 860:2007 e a NBR 13790:1997 acerca da harmonização de termos e conceitos, vale apresentar alguns trabalhos nessa área que abordam a aplicação da harmonização de termos e conceitos em domínios especializados. As dissertações *Harmonização da terminologia nos documentos produzidos no Gabinete de Intercâmbio do Ministério da Cultura de Angola* (Segunda, 2017) e *Variantes Terminológicas: necessidade de harmonização nos regulamentos internos do Ministério da Ciência e Tecnologia* (Lima, 2016), ambas produzidos na Universidade Nova Lisboa, abordam a aplicação da harmonização de termos em um *corpus* de textos jurídicos monolíngue, desaconselhando certos usos, recomendando termos mais frequentes e conferindo um certo grau de consistência entre os respectivos termos. Além desses, também encontramos a dissertação *Proposta de harmonização da terminologia designativa de área e subáreas do português como*

*língua não materna baseada em corpus* (Lisboa, 2021), produzida na Universidade Federal de Uberlândia, que propôs uma harmonização terminológica da área de Português como Língua Não Materna (PLNM) e com suporte de um *corpus* composto por textos da área investigada. Lisboa (2021) apresenta as relações entre os conceitos encontrados e evidencia os limites conceituais. Os trabalhos citados acima apontam para a possibilidade de se aplicar a harmonização terminológica em diferentes domínios, mas nenhum dos trabalhos aproximou *frames* e harmonização, foco deste estudo.

Além desses estudos, Mahr (2023), em seu artigo intitulado *The impact of ISO/TC 37 standards on technical communication*, discute sobre como as recomendações da ISO/TC 37 referentes aos trabalhos terminológicos impactam na comunicação técnica e na tradução. Mahr (2023) apresenta a padronização em casos que envolvem produtos de mercado global, os quais exigem um conhecimento específico dos empregados que entendem os conceitos do domínio. Nessas situações, percebe-se que há a necessidade de se encontrar termos apropriados em muitas línguas, ou até mesmo na mesma língua em que há variações culturais e sociais.

Em se tratando da padronização ou harmonização de termos e conceitos no contexto apresentado acima, Mahr (2023) salienta que se utilizam algumas abordagens mais cognitivas para tentar manipular a terminologia. Geralmente, recorre-se aos postulados da Psicologia Cognitiva, da Linguística e da Programação para que seja possível lidar com duas funcionalidades da comunicação técnica nessa situação: (i) gerar uma representação do conhecimento junto à informação técnica orientada para o uso de um determinado produto, servindo como uma estrutura para a troca de informações entre fornecedores, vendedores e usuários; e (ii) apresentar uma interoperabilidade com outras arquiteturas relacionadas ao produto, como, por exemplo, uma placa de identificação digital de um produto específico. Desse modo, a padronização e a harmonização se tornam fulcrais para a comunicação técnica ser eficiente e para a organização dos produtos.

Considerando as discussões acerca dos *frames* semânticos, da Terminologia baseada em *Frames* e da concepção e aplicação de harmonização de termos e conceitos, é importante ressaltar, inicialmente, que o conceito, a nosso ver, apresenta dois aspectos: o lexical e o conceptual. É justamente por isso que apostamos na contribuição dos *frames* para desenvolver esta nova metodologia semântico-terminológica. Assim, ao aproximarmos tais teorias, percebemos que é viável desenvolver uma abordagem metodológica que abranja *frames* para incrementar o processo de harmonização.



Quando passamos a considerar o aspecto contextual e a visão enciclopédica da Semântica de *Frames* nessa nova estrutura para o processo de harmonizar, ponto que não é desenvolvido pela ISO, é possível compreender melhor as relações entre conceitos e termos, o que facilitará, posteriormente, a harmonização. Vale enfatizar que compreender e analisar as estruturas de conceitos e termos, para a ISO, é uma das etapas fundamentais para aplicar a harmonização e, nesse sentido, os *frames* auxiliarão nesse processo. Outro ponto que é fundamental salientar diz respeito às duras críticas feitas aos documentos propostos pela ISO, que não atentam para os aspectos social, cognitivo e comunicativo na sua elaboração (Cabré, 1999; Temmerman, 2004; Faber, 2012). Isso estimula, também, o desenvolvimento da nossa proposta metodológica, uma vez que, como dito anteriormente, esses pontos serão considerados em nosso projeto.

Em consonância à afirmação de Faber (2012), as unidades lexicais no contexto especializado podem evocar *frames*, e isso se refletirá na construção das estruturas de *frames* do ERP, visto que ele integra diversas áreas que fazem parte de uma empresa. Além disso, quanto à afirmativa de Faber (2012), será possível analisar o comportamento das unidades lexicais, apresentando, inclusive, exemplos em uso. No que se refere à compreensão do comportamento das unidades lexicais tanto no contexto de cada departamento como no contexto maior, do ERP, será possível refinar a harmonização e arquitetar toda a estrutura semântico-terminológica do *software*. Para a aplicação automática da harmonização no ChatGPT, a elaboração dos *frames* e da estrutura do *Knowledge Graph* contribuirão para que o modelo de linguagem compreenda a terminologia e os contextos em que são utilizados. Essa camada de informações advinda dos *frames* é fundamental para que seja possível visualizar de forma mais abrangente os termos e selecionar qual será o termo principal e quais serão as variantes tanto na harmonização manual como na automática. Com relação a se utilizar exemplos em uso e inseridos em um contexto, essa possibilidade tornaria o produto mais completo e poderia, também, ser útil para os usuários da ferramenta. A partir dos exemplos, os clientes que utilizam o *software* seriam capazes de aproveitá-los para compreender melhor determinado termo, descobrir qual terminologia empregar em determinado contexto, ou como proceder em uma dada situação empresarial que exija seus conhecimentos técnico-terminológicos.

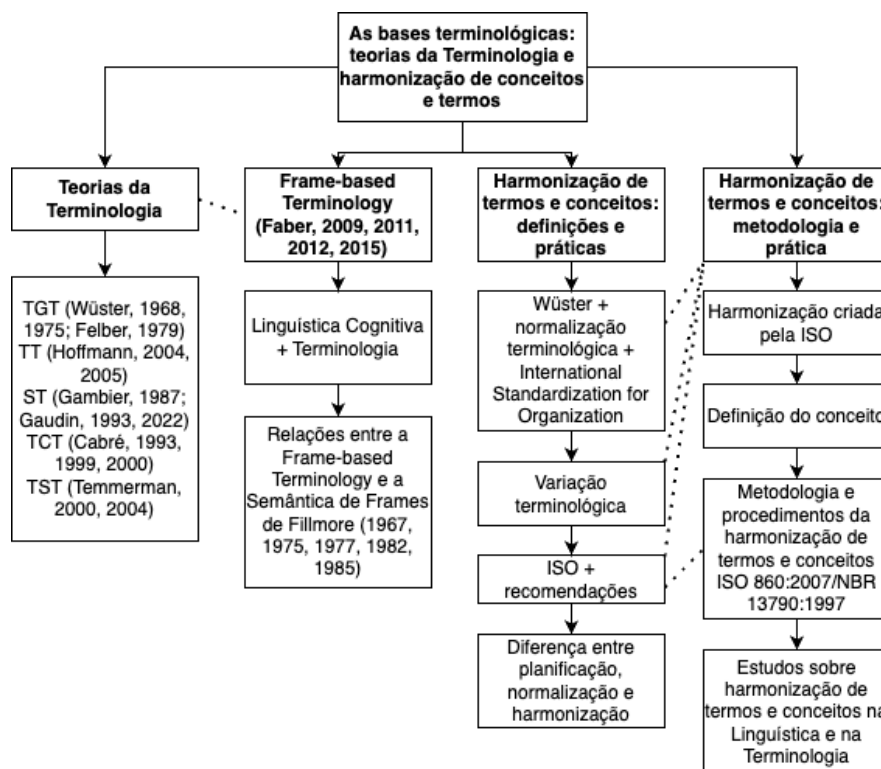
Tendo em vista possíveis contribuições para o desenvolvimento de nossa proposta de metodologia semântico-terminológica, não encontramos trabalhos que unam as teorias da Semântica e da Terminologia para reformular o processo de

harmonização de termos e conceitos. Durante nosso estudo, examinamos trabalhos que tratam da temática terminológica associada aos *frames* semânticos, como os recursos *DiCoEnviro*, *EcoLexicon*, *Field: dicionário de expressões do futebol*, *Dicionário Olímpico* e *Dicionário Paralímpico*, bem como trabalhos de aplicação da harmonização (cf. Lima, 2016; Segunda, 2017; Lisboa, 2021; Mahr, 2023). No entanto, nenhum dos estudos se dedicou a relacionar as áreas para o tratamento dos termos e conceitos, já que as ferramentas baseadas na Semântica e na Terminologia não aplicaram a harmonização e os trabalhos de harmonizar termos seguiram puramente o viés terminológico.

Por fim, ressaltamos que um dos diferenciais deste trabalho é a entrada de linguistas em trabalhos tecnológicos de empresas voltadas estritamente para a Computação. Tradicionalmente, a busca por soluções para aprimorar um *software* de gerenciamento empresarial seria feita apenas por profissionais da área computacional. Contudo, com a necessidade de resolver impasses de ordem semântico-terminológica, a aproximação entre a Linguística Aplicada e a Computação é fundamental para investigar e aprimorar tais demandas.

Considerando os pontos discutidos neste capítulo, apresentamos a Figura 24, com a intenção de indicar os conceitos e autores trabalhados.

Figura 24 - Síntese do capítulo “As bases terminológicas: teorias da Terminologia e harmonização de termos e conceitos”



Fonte: Elaborada pela autora.

Como ilustra a Figura 24, traçamos um histórico das teorias terminológicas, apresentando a Teoria Geral da Terminologia, a Terminologia Textual, a Socioterminologia, a Teoria Comunicativa da Terminologia e a Teoria Sociocognitiva da Terminologia. Em seguida, abordamos a *Frame-based Terminology*, teoria desenvolvida por Faber, que relaciona Terminologia, Linguística Cognitiva e Semântica de *Frames*. Posteriormente, tratamos das definições e práticas relacionadas à harmonização de termos e conceitos, discutindo sobre a influência de Wüster na concepção de diversas normas da *International Organization for Standardization (ISO)* e sobre a variação terminológica. Após isso, exploramos algumas recomendações da ISO, diferenciamos os conceitos de planificação, normalização e harmonização. Por fim, apresentamos a metodologia e a prática da harmonização de termos e conceitos, expondo algumas definições sobre termos e conceitos do ponto de vista da ISO e trabalhos relacionados a essa temática.

Após explorarmos as bases terminológicas deste trabalho, no próximo capítulo, A interface Linguístico-Computacional: Inteligência Artificial, Processamento de Linguagem Natural e *Large Language Model*, daremos atenção à abordagem

computacional deste trabalho, tratando da Inteligência Artificial, do Processamento de Linguagem Natural, dos *Large Language Models* (LLMs), da representação do conhecimento, como as ontologias, os *Knowledge Graphs* e os *frames*. Depois disso, apresentaremos trabalhos que tratam da harmonização de termos e conceitos pela perspectiva da Computação.

## **4 A INTERFACE LINGUÍSTICO-COMPUTACIONAL: INTELIGÊNCIA ARTIFICIAL, PROCESSAMENTO DE LINGUAGEM NATURAL E *LARGE LANGUAGE MODEL***

Este trabalho tem por objetivo, em nível teórico, investigar o papel da Semântica de *Frames* como princípio organizador de uma base de dados ERP, contribuindo para a harmonização terminológica e, em nível prático, desenvolver um *Knowledge Graph* com *frames* semânticos do ERP, bem como explorar estratégias e técnicas de *prompt* para trabalhar a harmonização de termos e conceitos utilizando o ChatGPT. Assim, neste capítulo, abordaremos a parte computacional e aplicada desta tese. Para tanto, na subseção 4.1 Inteligência Artificial e Processamento de Linguagem Natural: bases, técnicas e métodos, trataremos, inicialmente, da aproximação da Filosofia, Matemática, Psicologia e Linguística na IA para, em seguida, contar a história da área da Inteligência Artificial e, posteriormente, abordar o Processamento de Linguagem Natural (PLN), subárea da Ciência da Computação, da Inteligência Artificial e da Linguística. Na sequência, discorreremos sobre os Modelos de Linguagem e os Grandes Modelos de Linguagem (LMs e LLMs). Por fim, na subseção 4.2, discutiremos sobre a Representação do Conhecimento, explorando os conceitos de Ontologia, *Knowledge Graphs* e *Frames*.

### 4.1 Inteligência Artificial e Processamento de Linguagem Natural: bases, técnicas e métodos

Para tratar dos aspectos que abrangem a Inteligência Artificial (IA), é importante tratar dos fundamentos e métodos basilares da área, os quais se relacionam com as principais noções e abordagens advindas das áreas da Filosofia, Matemática, Psicologia e Linguística para, então, ocupar-se da história da Inteligência Artificial na Ciência da Computação. Tendo isso em mente, primeiramente, apresentaremos as discussões filosóficas, matemáticas, psicológicas e linguísticas que fundamentam a IA e, depois, contaremos a história da IA pela perspectiva do campo computacional.

#### 4.1.1 Filosofia, Matemática, Psicologia e Linguística e relações com IA e PLN

Desvendar os mistérios que envolvem a inteligência sempre moveu o homem. Tentar entender como agimos, pensamos e como o cérebro pode perceber, compreender, prever e manipular um mundo tão complexo ainda é um desafio. Na Ciência da Computação, uma das áreas que foca em estudos e no desenvolvimento de agentes inteligentes é a Inteligência Artificial (IA), que se ocupa da automação do comportamento inteligente e, como destaca Luger (2013), deve seguir princípios teóricos e aplicados bem fundamentados. Tais princípios devem considerar as estruturas de dados utilizados na representação do conhecimento, os algoritmos necessários para aplicar o conhecimento e as linguagens e técnicas de programação usadas em sua implementação (Luger, 2013). Além de se preocupar com a compreensão da inteligência, Russell e Norvig (2022) salientam que a IA também busca construir entidades inteligentes, ou seja, máquinas que possam processar e agir de forma eficiente e segura em diferentes situações e tarefas intelectuais. Atualmente, as aplicações desenvolvidas no âmbito de IA abarcam diversas possibilidades, desde as mais gerais, como raciocinar, aprender e perceber as coisas, até tarefas mais específicas como jogar, escrever, produzir conteúdo, dirigir sozinho e diagnosticar doenças (Lima; Pinheiro; Santos, 2014).

No curso da história da IA, Luger (2013) e Russell e Norvig (2022) afirmam que os questionamentos advindos de diferentes áreas são cruciais para a fundamentação das teorias e dos métodos na Inteligência Artificial. Aqui, destacamos os pressupostos da Filosofia, da Matemática, da Psicologia e da Linguística para apresentar as bases da IA.

No campo da Filosofia, o primeiro a tratar das questões do homem e da mente, foi Aristóteles quem abordou a concepção do pensamento correto (o processo de raciocínio irrefutável), segundo a qual as leis do pensamento tinham como função governar a parte racional da mente (Luger, 2021). Assim, criou-se um sistema informal de silogismos para um raciocínio adequado, que possibilitava gerar conclusões de forma mecânica, considerando as premissas dadas inicialmente. Além disso, o filósofo discutiu sobre a conexão entre conhecimento e ação. De acordo com Russell e Norvig (2022), Aristóteles sustentou a visão de que as ações são justificadas por uma conexão lógica entre os objetivos e o conhecimento do resultado da ação. Para lidar com as ações e o raciocínio, nessa perspectiva, Aristóteles sugeriu a criação de um algoritmo capaz de lidar com tal concepção.

Outro ponto discutido por Aristóteles é a distinção entre matéria e forma, que serviu de base para as noções modernas da Computação Simbólica<sup>25</sup> e a abstração de dados. À luz dos pressupostos aristotélicos, a mente podia ser vista (i) a partir do seu funcionamento que, em partes, seguia regras lógicas e numéricas (além de buscar a construção de sistemas físicos que levavam em conta algumas dessas regras) e (ii) como um próprio sistema físico (Luger, 2021; Russell; Norvig, 2022).

Mais adiante, Descartes retoma a discussão acerca da distinção entre mente e matéria, apresentando o Dualismo. Para ele, uma parte da mente humana, considerada como fora da natureza, é isenta das leis físicas. Logo, os animais, por exemplo, que não possuem essa qualidade dualista, poderiam ser considerados como máquinas (Russell; Norvig, 2022). Um caminho alternativo para o Dualismo é o Materialismo, que compreende que o funcionamento do cérebro segue as leis da Física e são elas que constituem a mente. Desse modo, é a mente física que manipula o conhecimento; todavia, a problemática dessa perspectiva está em como estabelecer a fonte do conhecimento.

As contribuições filosóficas de Aristóteles e Descartes (dentre outros filósofos) ligam-se à corrente racionalista, que se apoiava na visão lógica e formalizada. Contrário aos postulados do racionalismo, o movimento empirista fundamenta-se na máxima de John Locke: nada entra na mente senão por meio dos sentidos (Luger, 2013). Sob essa perspectiva, passa-se a distinguir dois tipos de fenômenos mentais: (i) a percepção e (ii) o pensamento, a memória e a imaginação. Por esse viés, Russell e Norvig (2022) apresentam o Princípio de Indução, desenvolvido por David Hume, em que as regras gerais são adquiridas a partir da exposição às associações repetidas entre seus elementos. Conforme elucida Luger (2013), determinadas características perceptivas são associadas por meio de experiências repetitivas, criando uma disposição na mente para relacionar as ideias correspondentes.

Posteriormente, conforme aponta Luger (2013; 2021) Ludwig Wittgenstein passa a reconsiderar a base do significado – questão não examinada no racionalismo e na visão lógica – tanto no que tange à linguagem natural quanto à formal. Logo, o significado de qualquer expressão vocal depende dos contextos humano e cultural em que ela está situada. Nesse sentido, Wittgenstein passa a questionar os pressupostos

---

<sup>25</sup>A computação simbólica é “baseada na hipótese de um sistema de símbolos definido, em que há um conjunto de estruturas simbólicas e um conjunto de regras de manipulação dessas estruturas, pelos quais têm-se os meios necessários e suficientes para se criar comportamento inteligente” (Lima; Pinheiro; Santos, 2014, [p. 5]).

racionalistas atinentes aos fundamentos da linguagem, do conhecimento e da ciência. O filósofo argumenta que é preciso considerar o uso da linguagem pensando nas escolhas feitas e nas ações realizadas dado um contexto cultural variável. Para ele, assim como o uso da linguagem, a ciência e a matemática também são construções sociais e, portanto, devem considerar o contexto. De acordo com Skansi (2018), as concepções filosóficas de Wittgenstein, especialmente no que se referem à reflexão sobre o contexto nos estudos da linguagem, desempenharam um papel fundamental na história do Processamento de Linguagem Natural e serviram de base para o desenvolvimento de abordagens, como, por exemplo, a dos *embeddings*.

Como destacam Russell e Norvig (2022), mesmo com todas as contribuições que fundamentaram a Inteligência Artificial, para ser considerada como ciência formal, foi preciso matematizar a lógica e a probabilidade e introduzir um novo ramo nas Ciências Matemáticas: a Computação. Nesse contexto, o pensamento era visto como uma forma de cálculo formalizado e mecanizado. A concepção da lógica formal da Filosofia começou realmente a ser desenvolvida pelo matemático George Boole. Em seus trabalhos, Boole dedicou-se à formalização das leis da lógica e aos detalhes da lógica proposicional (Skansi, 2018). Tais pontos representam o cerne da Ciência da Computação moderna.

Mais tarde, em 1879, Gottlob Frege formalizou muitas questões tratadas por Aristóteles e expandiu a lógica booleana para adicionar objetos e relações, criando a Lógica de Primeira Ordem (Skansi, 2018; Russell; Norvig, 2022). De acordo com Luger (2013), tal formalização de Frege resultou na linguagem conhecida como Cálculo de Predicados de Primeira Ordem, a qual inclui símbolos de predicados, teoria de funções e variáveis quantificadas, sendo engendrada como uma linguagem que descreve a matemática e suas bases filosóficas. Além de Frege, Russell também tratou de formalizar a linguagem, mostrando que os axiomas e os teoremas seriam tratados apenas como sequências de caracteres e, para as provas (resultado de aplicações de regras sintáticas), seriam processadas apenas pela aplicação de regras bem definidas capazes de manipular as sequências (Russell; Norvig, 2022). Pela perspectiva logicista, Frege e Russell buscavam formalizar o pensamento lógico ou matemático por meio de símbolos. Nesse sentido, acreditava-se que seria possível gerar uma linguagem artificial sem ambiguidade e conferir-lhe um certo grau de certeza, já que era baseada em regras lógicas na sua manipulação. Dessa forma, a crença de que “pensar é calcular” era possível justamente pela formalização simbólica e pelo desenvolvimento de uma



máquina que fosse capaz de executar tal linguagem (Correa, 2006; Chiavegatto, 2009). Contudo, vale ressaltar que a grande ameaça a essa visão era a possibilidade de que alguns aspectos da cognição não poderiam ser formalizados como se pensava.

Ainda no campo da lógica, a Teoria da probabilidade foi desenvolvida como uma lógica generalizante para situações com informações incertas. Os estudos acerca da probabilidade permitiram que as ciências quantitativas pudessem lidar com medições incertas e teorias incompletas. Com a formalização da probabilidade e os dados disponíveis, houve a expansão da estatística (Russell; Norvig, 2022).

Assim como a Filosofia e a Matemática, a Psicologia também embasou a área da Inteligência Artificial. Os primeiros trabalhos desenvolvidos na Psicologia tratavam da aplicação de um método científico para estudar a visão humana. Como os experimentos desenvolvidos nessa área não apresentavam a natureza objetiva dos dados e o pensamento era introspectivo, biólogos, que buscavam compreender o comportamento humano a partir da análise do comportamento animal, não puderam utilizar as bases da Psicologia para seus estudos e, assim, desenvolveram uma metodologia mais objetiva (Luger, 2013; Russell; Norvig, 2022). Seguindo a metodologia elaborada na Biologia para compreender o comportamento humano, o movimento Behaviorista, liderado por John Watson, rejeitava qualquer estudo que abrangesse apenas os processos mentais, argumentando que a introspecção não poderia fornecer dados confiáveis (Russell; Norvig, 2022). Logo, o behaviorismo investigava somente os estímulos dados a um animal e suas respostas ao estímulo fornecido.

Segundo Skansi (2018) e Luger (2021), posteriormente, a Psicologia Cognitiva passa a compreender o cérebro como um dispositivo que processa informações. Nos Estados Unidos, a modelagem computacional baseada nessa abordagem levou à criação do campo das Ciências Cognitivas. Nessa perspectiva, o cognitivismo trabalhava com a ideia de que a mente, que era vista como um sistema complexo e composto por muitas partes que interagem entre si, deveria ser examinada por conta própria, isto é, ela é independente do cérebro, e os estudos deveriam seguir métodos formais para a sua análise. Nesse sentido, buscava-se criar e analisar sistemas que tentassem recriar partes da realidade neurológica e, concomitantemente, deveriam ser capazes de recriar o comportamento.

Baseados nos pressupostos cognitivos, quatro teóricos das áreas da psicologia, linguística e computação, George Miller, Noam Chomsky e Allen Newell e Herbert Simon, respectivamente, uniram-se para mostrar como modelos de computador

poderiam ser utilizados para abordar a psicologia da memória, da linguagem e do pensamento lógico. Considerando as teorias de seis disciplinas diferentes, Antropologia, Ciência da Computação, Linguística, Neurociência, Filosofia e Psicologia, as Ciências Cognitivas aproximaram o conceito de mente para o interior do campo científico. Sob essa ótica, separava-se mente e corpo, distinguindo o que era processo interno do que era processo externo.

No âmbito da Linguística, a quebra de paradigma behaviorista e a virada cognitiva ocorreram, sobretudo, com a Gramática Universal de Chomsky e os ataques contra o Behaviorismo de Skinner (Skansi, 2018). Apoiado pela psicologia behaviorista, Skinner aproximou os pressupostos do behaviorismo à aprendizagem de línguas. Contrário a essa visão, Noam Chomsky defende que tal corrente não considerava a criatividade da língua e aponta que ela não poderia explicar, por exemplo, como as crianças podiam entender e criar frases que nunca tinham ouvido antes, ou como os bebês balbuciavam (Luger, 2013). Nesse contexto, nasce a Linguística Gerativa, teoria baseada em modelos sintáticos, extremamente formal, com o intuito de poder ser programada. Por esse viés, a representação da linguagem era vista como símbolos, e o processamento mental como manipulação desses símbolos; logo, associavam-se a lógica da matemática e as estruturas linguísticas. No cognitivismo, defende-se que uma teoria cognitiva deveria ser como um programa de computador, isto é, considerar a cognição e a linguagem humana em todos os seus aspectos (ainda que muitos desses aspectos tenham sido deixados de lado) e poder formalizá-las para que fossem programadas (Russell; Norvig, 2022). Posteriormente, abordaremos de forma aprofundada as questões da Linguística Gerativa quando tratarmos do campo de Processamento de Linguagem Natural.

A Linguística, nesse período, herdou os pressupostos logicistas e formais. No entanto, a semântica não era vista na sua completude. Ressalta-se que, nessa perspectiva de formalizar a linguagem, a única ameaça visível era a possibilidade de que alguns aspectos da cognição não pudessem ser formalizados. Consequentemente, três pontos desestruturaram o cognitivismo clássico: (i) a computação não é, necessariamente, simbólica; (ii) mente e corpo não são entidades estanques; e (iii) as atividades cognitivas não são separadas do meio, isto é, compreendemos o mundo e a linguagem por meio das experiências que vivemos. Desse modo, abre-se espaço para refletir acerca dos aspectos sociais e culturais, que são intrínsecos à língua, e as questões semânticas passam a ganhar espaço nos estudos linguísticos e computacionais.

Dando continuidade a esta discussão, na sequência, trazemos alguns dos principais conceitos e abordagens das diferentes áreas, fundamentais para compreender as bases da Inteligência Artificial, começando pela história da IA.

#### 4.1.2 A história da Inteligência Artificial

O surgimento da Inteligência Artificial como campo de estudo ocorreu em 1956, dois meses após o *Dartmouth Summer Research Project on Artificial Intelligence*, proposto por John McCarthy, Marvin Minsky, Claude Shannon e Nathaniel Rochester (Lima; Pinheiro; Santos, 2014; Skansi, 2018; Luger, 2021). Contudo, alguns trabalhos anteriores a 56 já haviam sido elaborados. O primeiro estudo reconhecido no campo da Inteligência Artificial foi desenvolvido por Warren McCulloch e Walter Pitts, em 1943, os quais propuseram um modelo de neurônios artificiais, que “ligavam” e “desligavam” ao reagirem a estímulos por um número suficiente de neurônios vizinhos (Lima; Pinheiro; Santos, 2014). Para esse modelo, os autores combinaram a modelagem matemática, o conhecimento de fisiologia básica e da função dos neurônios no cérebro, a análise formal da lógica proposicional e a teoria da computação desenvolvida por Alan Turing (Russell; Norvig, 2022).

Alan Turing, considerado o pai da Computação, atuou como matemático e cientista da computação. Durante a Segunda Guerra Mundial, Turing trabalhou para a Inteligência britânica com o objetivo de decifrar as mensagens nazistas geradas pela máquina de criptografia ‘Enigma’. Para essa tarefa, o matemático desenvolveu um sistema capaz de traduzir os textos de forma que eles fossem legíveis e compreensíveis (Frana; Klein, 2021; Taulli, 2019; Lima; Pinheiro; Santos, 2014; Skansi, 2018; Luger, 2021). Posteriormente, em 1947, ele já tratava sobre Inteligência Artificial e, em 1950, publicou um artigo debatendo sobre o funcionamento de computadores e se eles se igualariam ao desempenho humano. Nesse artigo, Turing também propôs um teste para fornecer uma definição operacional de inteligência. O teste consistia em um homem, uma mulher e um juiz (que poderia ser tanto um homem como uma mulher), em salas distintas, que deveriam se comunicar somente por meio de textos. O objetivo era que o homem se passasse por mulher, e vice-versa, enganando o juiz. Na próxima etapa, Turing substituíria um deles por um computador. Logo, ainda em salas distintas, o computador seria interrogado por um humano em um terminal, e o sistema deveria responder como se fosse um ser humano. Assim, o juiz analisaria o conteúdo da

conversa e diria se havia um computador ou um indivíduo do outro lado da linha (Rosa, 2011; Onody, 2021). Conforme afirma Rosa (2011), para o teste ser satisfatório e apresentar o comportamento humano, a máquina deveria desenvolver a habilidade de atingir a performance humana em todas as tarefas cognitivas.

De acordo com Russell e Norvig (2022) e Skansi (2018), em 1950, Marvin Minsky e Dean Edmonds construíram o primeiro computador de rede neural. Em 1955, John McCarthy, Marvin Minsky, Claude Shannon e Nathaniel Rochester reuniram investigadores americanos que se interessavam pela teoria dos autômatos, pelas redes neurais e pelos estudos de Inteligência. A conferência foi realizada em 1956, com dez participantes. Dentre eles destacam-se Allen Newell e Herbert Simon, que apresentaram a “*Logic Theorist*”, um programa computacional capaz de pensar de forma não numérica, resolvendo o problema mente-corpo.

A década de 50 foi marcada pela crença de que uma máquina jamais poderia fazer algo como o homem. Tal convicção motivou os pesquisadores na área de IA a se concentrarem e desenvolveram tarefas indicativas de inteligência, como jogar, montar quebra-cabeças, resolver problemas matemáticos e realizar testes de QI. Nessa época, Russell e Norvig (2022) e Skansi (2018) destacam o trabalho desenvolvido por Allen Newell e Herbert Simon, que continuaram com a “*Logic Theorist*” e criaram o *General Problem Solver* (GPS), um programa que buscava imitar protocolos humanos de resolução de problemas. Além disso, Arthur Simon desenvolveu o jogo de damas, utilizando métodos que hoje conhecemos como aprendizagem por reforço, e o programa aprendeu a jogar em nível amador. Com isso, Simon refutou a ideia de que os computadores só realizavam o que lhes era mandado, visto que o programa aprendeu a jogar mais rápido e melhor do que o seu criador (Luger, 2021). Ainda na década de 50, John McCarthy apresentou duas grandes contribuições para a área da Inteligência Artificial. A primeira contribuição foi a definição de uma linguagem de alto nível, LISP, que passou a ser utilizada como a linguagem de programação de IA; a segunda foi a apresentação de uma proposta conceitual para sistemas de IA baseadas em conhecimento e raciocínio. Com sua proposta, McCarthy introduziu alguns princípios centrais da representação do conhecimento ao integrar o conhecimento de mundo e usá-lo para elaborar planos de ação.

Os desenvolvimentos computacionais que ocorreram nos anos 50 motivaram os pesquisadores da área de Inteligência Artificial e deram mais confiança à criação de sistemas e programas. Contudo, com a convicção de que o sucesso da IA estaria muito

próximo, como, por exemplo, desenvolver computadores capazes de serem campeões em xadrez ou de resolver um teorema matemático, muitos projetos fracassaram e, mais tarde, foram novamente realizados e provados muito tempo depois. Russell e Norvig (2022) apontam duas razões principais para o fracasso de tais sistemas. A primeira diz respeito aos primeiros sistemas que se fundamentavam, principalmente, na introspecção de como o homem executava determinada tarefa e não seguia uma análise cuidadosa da tarefa em si. A segunda, por sua vez, se refere à ausência de uma apreciação que justificasse a falta de solução de muitos problemas que a IA buscava resolver. Nesse sentido, muitos investigadores acreditavam que o problema dos sistemas estava ligado à falta de um *hardware* rápido e com memória suficiente; entretanto, o empecilho estava na própria teoria.

Os primeiros experimentos e as primeiras pesquisas em IA são conhecidos como Abordagens por Método Fraco (Luger, 2013; Russell; Norvig, 2022). Esse enfoque é marcado por programas que se restringiram a estratégias baseadas em sintaxe e objetivaram criar diferentes aplicações. Embora as pesquisas estivessem voltadas para a elaboração de mecanismos de buscas que encadeavam etapas de raciocínio para soluções completas, elas não se adaptavam à resolução de problemas mais complexos. Pensando nisso, uma alternativa era explorar o uso de conhecimentos mais poderosos e específicos de domínio, que permitiam etapas de raciocínios maiores e que pudessem lidar com situações específicas de áreas de especialização. Essa nova perspectiva impactou também nos estudos na área de compreensão de linguagem natural, visto que muitos problemas linguísticos, como a ambiguidade, não eram solucionados e necessitavam de uma compreensão mais robusta da língua e do desenvolvimento de métodos mais eficazes para utilizar o conhecimento. Essa nova visão ficou conhecida como Abordagem por Método Forte, já que utilizava o conhecimento explícito de um domínio de problema particular (Luger, 2013; Lima; Pinheiro; Santos, 2014; Skansi, 2018).

Nesse contexto, houve um crescimento significativo na construção de aplicações para problemas do mundo real e, com isso, muitas ferramentas de representação e raciocínio foram elaboradas. Alguns recursos ainda seguiam a perspectiva lógica, como a linguagem Prolog; outros, como os *frames* de Minsky<sup>26</sup>, adotaram uma abordagem mais estruturada, apresentando um esquema representacional que capturava as conexões

---

<sup>26</sup> Na subseção 4.2, trataremos da representação do conhecimento, abordando os *frames* de Minsky de forma mais aprofundada.

implícitas da informação em um domínio de problema (Luger, 2013). Com as diferentes perspectivas na construção de programas e sistemas, de 1980 a 1988, a indústria de IA cresceu de forma exponencial, centenas de empresas passaram a desenvolver sistemas especializados, sistemas de visão, robôs, *softwares* e *hardwares* (Russell; Norvig, 2022).

Posteriormente, com a fragilidade dos sistemas especialistas, uma nova abordagem foi adotada. Tal enfoque incorporava probabilidade, aprendizagem automática e resultados experimentais em vez de lógica booleana, codificação manual e afirmações filosóficas. Assim, surgiu o raciocínio probabilístico e o *Machine Learning*, que utilizavam teorias matemáticas rigorosas e os modelos eram gerados por processo de treinamento em um grande conjunto de dados reais de fala. Judea Pearl, em 1988, aplicou o raciocínio probabilístico em sistemas inteligentes. As redes bayesianas elaboradas por Pearl produziram um formalismo rigoroso e eficiente para a representação do conhecimento incerto e algoritmos práticos para raciocínio probabilístico (Russell; Norvig, 2022). Com essa nova abordagem de IA, que reconhece a importância dos dados, da modelagem estatística e da otimização e da aprendizagem de máquina, subcampos da área se reaproximaram, como a visão computacional, a robótica, o reconhecimento de fala, os sistemas multiagentes e o Processamento de Linguagem Natural.

Com os avanços tecnológicos e a WWW (*World Wide Web*), percebeu-se a facilidade em criar conjuntos de dados enormes, podendo incluir trilhões de palavras de texto, bilhões de imagens, bilhões de horas de fala e vídeo, dados de rastreamento, dados de redes sociais, dentre outras possibilidades. Essa nova era da Inteligência Artificial é marcada pelo *Big Data*. A disponibilidade de grandes conjuntos de dados e a mudança para o *Machine Learning* contribuíram para que a atração comercial pela IA voltasse a crescer. De acordo com Luger (2021) e Russell e Norvig (2022), o *Big Data*, por exemplo, foi essencial para que o sistema Watson da IBM vencesse o campeão humano no programa *Jeopardy* (programa televisivo de perguntas e respostas).

Até a segunda metade do século XXI, observou-se um avanço significativo no desenvolvimento de modelos de linguagem<sup>27</sup> baseados em representação numérica, seguindo os preceitos da IA Simbólica. Nesse sentido, a linguagem era formalizada em

---

<sup>27</sup> Um modelo de linguagem é um sistema de Inteligência artificial e um dos principais componentes de Processamento de Linguagem Natural, uma vez que permitem que as máquinas processem e gerem linguagem natural. O modelo de linguagem, geralmente, é treinado com um grande conjunto de dados para aprender a prever a próxima palavra de uma sequência com base no contexto.

formato matemático, como códigos e até mesmo codificações, e geravam e utilizavam textos mapeados para representação numérica (Paes; Vianna; Rodrigues, 2023). Tal representação facilitava a implementação no computador e poderia auxiliar na representação por indexação. Contudo, representar os componentes linguísticos de forma numérica não compreendia o aspecto semântico da língua. De acordo com Paes, Vianna e Rodrigues (2023, p. 2),

Um modelo é uma simplificação de um fenômeno complexo, no nosso caso, uma simplificação da língua que possa ser representada por ferramentas computacionais. Embora um modelo tente capturar as nuances do fenômeno real, justamente por ser uma simplificação, ele não tem a intenção de substituir o fenômeno real, mas representá-lo para auxiliar o nosso entendimento ou resolver algumas tarefas.

Assim, um modelo de linguagem deveria considerar todos os princípios linguísticos, como o lexical, sintático e semântico. Tendo em vista que o atributo semântico não era representado, uma solução foi mapear os componentes da língua por meio de vetores em um dado espaço semântico (*embeddings*), conforme a hipótese distribucional (Skansi, 2018). Para Paes, Vianna e Rodrigues (2023), os modelos semânticos distribucionais consideram o contexto a partir da coocorrência de itens. Dessa forma, as palavras que aparecem próximas de outras apresentam contextos similares.

Em um primeiro momento, um modelo de linguagem deveria ser capaz de prever a próxima palavra da sequência a partir das palavras anteriores. Os modelos de linguagem probabilísticos, por exemplo, buscam prever as palavras em uma sequência com base em uma atribuição de probabilidade condicional das palavras (Skansi, 2018; Hagiwara, 2021). A probabilidade é feita a partir de dados advindos de um *corpus*. Quanto maior o conjunto de dados, maiores são as chances de apresentar variações de sequências (PAES; Vianna; Rodrigues, 2023). Embora o modelo probabilístico utilize *corpus*, contexto e volumosa quantidade de dados, dois pontos podem emergir: (i) o tempo de processamento é mais demorado devido à quantidade de dados e (ii) com a probabilidade, perde-se o aspecto da similaridade semântica. Uma solução para a questão da similaridade semântica estática foi utilizar modelos de linguagem neurais, capazes de gerar e aplicar *embeddings* contextualizados de forma dinâmica e vetorial (Skansi, 2018). Para nos aprofundarmos nesse tópico, é importante, primeiro, tratar de *Deep Learning*.

A possibilidade de acessar grandes quantidades de informações *online* e obter conjuntos de dados cada vez maiores permitiu que novas técnicas e novos métodos fossem desenvolvidos. Na era do *Machine Learning*, as máquinas ainda não possuíam processadores rápidos e eficientes, tornando a aprendizagem de máquina muito demorada. Com o crescimento do poder computacional, o *Deep Learning* apresentou uma nova forma de lidar com modelos e aplicações. Embora as redes neurais já tivessem sido trabalhadas nos anos 50 e 70 e no formato de redes neurais convolucionais em 1990 (Skansi, 2018; Paes; Vianna; Rodrigues, 2023), foi somente em 2011 que os métodos de *Deep Learning* impulsionaram: inicialmente, no campo de reconhecimento de fala e, depois, no reconhecimento visual de objetos. Sob esse viés, não se busca desenvolver computadores que pensem como seres humano, mas com um conjunto de grande volume de dados, processadores eficientes e rápidos e algoritmos sofisticados. “Os computadores podem começar a realizar tarefas que até então só podiam ser realizadas apenas por seres humanos, como reconhecer imagens e voz, criar obras de arte ou tomar decisões por si mesmo” (*Data Science Academy*<sup>28</sup>, 2022, local. cap. 3).

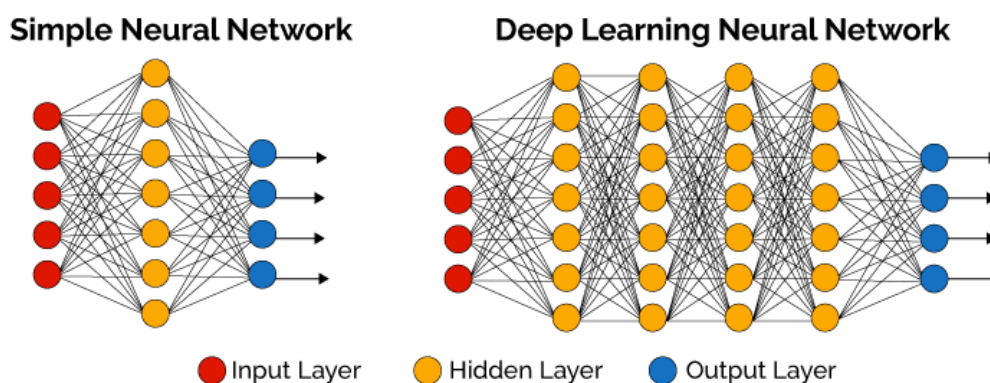
O *Deep Learning*, também chamado de Aprendizagem Profunda, é um subcampo do *Machine Learning*, que utiliza algoritmos para processar dados e reproduzir o processamento feito pelo cérebro do ser humano. Nesse sentido, o aprendizado de máquina utiliza múltiplas camadas de elementos da computação simples e ajustáveis (Russell; Norvig, 2022). As camadas utilizadas nessa abordagem são vistas como neurônios matemáticos que processam dados, compreendem a fala humana e reconhecem objetos (*Data Science Academy*, 2022). Assim, a informação é passada através de cada camada, como podemos observar na figura abaixo.

---

<sup>28</sup> *Data Science Academy*. O que são redes neurais profundas ou *Deep Learning*?. 2022. *E-book*. Disponível em: <https://www.deeplearningbook.com.br/o-que-sao-redes-neurais-artificiais-profundas/>. Acesso em: 14 jan. 2024.



Figura 25 - Rede neural simples e Rede neural profunda



Fonte: *Data Science Academy* (2022, local. cap. 3).

Na Figura 25, à esquerda, podemos ver a estrutura de uma rede neural simples (*Simple Neural Network*), um modelo matemático com a camada de entrada (*input layer* – em vermelho), uma camada escondida (*hidden layer* – em amarelo) e uma camada de saída (*output layer* – em azul). Ao lado, à direita, encontramos a rede neural profunda (*Deep Learning Neural Networks*), com uma camada de entrada (*input*), diversas camadas escondidas (*hidden layers*) e, finalmente, a camada de saída (*output*). As camadas escondidas, também chamadas de camadas ocultas, são algoritmos simples e semelhantes, contendo um tipo de função de ativação (*Data Science Academy, 2022*). As múltiplas camadas, nesse tipo de arquitetura, permitem que as redes neurais aprendam a resolver problemas mais complexos atinentes ao reconhecimento de padrões. Vale destacar que, para todas as camadas, é necessário atribuir pesos numéricos a fim de definir o parâmetro do neurônio. Os valores são determinados de acordo com o grau de evidência para a tomada de decisão. Além disso, as camadas de entrada e saída são diretas, ou seja, apresentam apenas um neurônio (diferenciando-se das camadas ocultas). Por serem diretas, não permitem que haja *looping*, etapa em que a rede de entrada recebe *feedback* da rede de saída e, conseqüentemente, estimula os demais neurônios. Caso a rede neural permita o *loop*, ela é chamada de Rede Neural Recorrente (*Data Science Academy, 2020*).

Voltando à similaridade semântica que discutimos anteriormente, os *embedding* estáticos não representavam dinamicamente as palavras no espaço semântico, bem como afetavam a ordem das palavras semanticamente distintas, mas lexicalmente idênticas (Paes; Vianna; Rodrigues, 2023). Com os métodos de Redes Neurais

Recorrentes e *Transformers*, é possível gerar e aplicar *embedding* contextualizados<sup>29</sup>. Atualmente, os métodos e as técnicas de *Deep Learning* permitem construir sistemas inteligentes, quase que autônomos. Um exemplo são os *Transformers*, modelos avançados de *Deep Learning*, introduzidos pelo departamento de pesquisa da Google, em 2017, capazes de lidar com diferentes tarefas de Processamento de Linguagem Natural tanto em pesquisas e experimentos acadêmicos como em desenvolvimento tecnológico da indústria (Ozdemir, 2023).

Como vimos, no curso da história da Inteligência Artificial, manipular dados e buscar métodos para representar o conhecimento e modelar a linguagem são cruciais para o desenvolvimento de sistemas inteligentes. Um campo que se destaca nesse contexto é o Processamento de Linguagem Natural, que será apresentado a seguir.

#### 4.1.3 Processamento de Linguagem Natural

O que distingue o ser humano das outras espécies é a forma como nos comunicamos. É por meio da linguagem que somos capazes de expressar nossos pensamentos e conhecimentos (Chen, 2021; Deng; Liu, 2018). A linguagem permite transmitir tais conhecimentos recorrendo a regras de linguagem e a diferentes componentes, incluindo os aspectos pragmáticos, semânticos, sintáticos e morfológicos. Construir significado e compreender o que está sendo produzido e dito não é uma tarefa humana simples. Para realizar essa mesma atividade usando um computador, é preciso converter a linguagem para que a máquina consiga interpretá-la e gerá-la (Singh, 2023), visto que não é apenas uma transmissão de palavras, mas também exige captar as inferências sobre o que se quer dizer, o conhecimento e as suposições do locutor, assim como o contexto da interação (Luger, 2013). Esse processo de pesquisa e modelagem da língua está ligado ao Processamento de Linguagem Natural (PLN).

De acordo com Kublik e Saboo (2022) e Caseli, Nunes e Pagano (2023), o Processamento de Linguagem Natural está ligado à área da Inteligência Artificial e se relaciona diretamente com a Linguística Computacional. Rosa (2011) afirma que o PLN apresenta diferentes definições, porém todas englobam a noção de manipulação de dados linguísticos e armazenamento destes em computadores. Logo, o objetivo

---

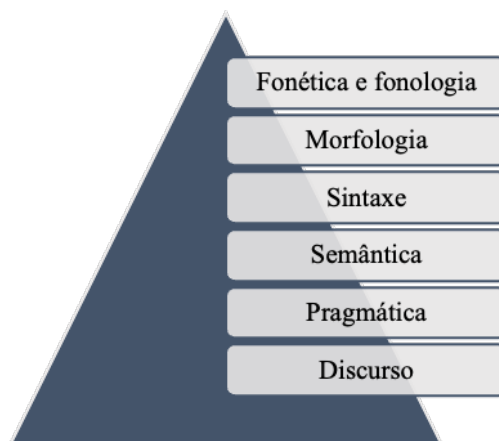
<sup>29</sup> Na subseção 4.1.3, trataremos do Processamento de Linguagem Natural, abordando o conceito de *embeddings*.

principal dessa área é investigar e desenvolver “métodos e sistemas de processamento computacional da linguagem humana” (Caseli; Nunes; Pagano, 2023, p. 1). Assim, as pesquisas no campo de PLN buscam trabalhar com teorias e metodologias que estudam o uso de computador para compreender e gerar de forma automática as linguagens humanas. É importante destacar que o ‘natural’ em PLN se refere à linguagem utilizada pelo homem no dia a dia para se comunicar, seja ela falada ou escrita; diferentemente, por exemplo, das linguagens de sinais matemática e artificial (linguagem de programação) (Zhang; Teng, 2021; Chen, 2021).

Em razão da heterogeneidade da linguagem natural, uma vez que usamos palavras diferentes para falar sobre coisas diferentes, e cada pessoa se expressa de uma forma, mesmo que tenhamos regras básicas em diferentes línguas, alguns aspectos como a variedade de estilos de escrita, as escolhas de palavras, as diferentes interpretações e a ambiguidade tornam as tarefas de Processamento de Linguagem Natural complexas e desafiadoras (Singh, 2023; Hagiwara, 2021). Logo, não podemos considerar somente o significado das palavras no dicionário. Luger (2013) destaca que é preciso desenvolver um processo complexo para capturar padrões de palavras, analisar sentenças, construir a representação do significado semântico e interpretá-lo tendo em vista o conhecimento do problema a ser investigado. Dessa forma, muitos trabalhos na interface entre Linguística e Computação investigam métodos computacionais para modelar linguagens e tratar de questões linguísticas (Zhang; Teng, 2021).

No campo de Processamento de Linguagem Natural, Caseli, Nunes e Pagano (2023) apontam duas grandes áreas de atuação: (i) a área de Interpretação (ou compreensão) de Linguagem Natural, que se dedica à análise e à interpretação da língua, à segmentação e à classificação de dados linguísticos e à interpretação e ao aprendizado de significado construídos pelo homem; e (ii) a área de geração de Linguagem Natural, a qual busca gerar respostas ao usuário de *chatbot*, assistentes virtuais como o ChatGPT e o Google Bard, por exemplo. Além das duas áreas de atuação, os estudos da linguagem se dividem em subáreas, conforme podemos observar na Figura 26.

Figura 26 - Subáreas dos estudos da linguagem



Fonte: Adaptada de Caseli, Nunes e Pagano (2023, p. 4).

No topo da pirâmide, encontramos o campo da fonética e da fonologia, que se dedica a estudar os sons e fonemas das palavras, área de grande importância para os estudos voltados para o reconhecimento e a geração computadorizada da fala. Em seguida, encontra-se a morfologia, que examina como morfemas se constituem em palavras. No terceiro nível está a sintaxe, responsável por investigar como as palavras se organizam e se estruturam, estudando as regras utilizadas para combinar palavras em frases e sentenças. Após a sintaxe, encontramos a área da semântica, que se ocupa do estudo do significado das palavras e frases. A pragmática, por sua vez, busca investigar a interação para fins comunicativos e, por fim, o discurso é a área que analisa o texto como um todo (Caseli; Nunes; Pagano, 2023). Além das seis subáreas apresentadas por Caseli, Nunes e Pagano (2023), Luger (2013) acrescenta mais dois níveis de análise para PLN: a prosódia, que se ocupa dos estudos ligados ao ritmo e à entonação da linguagem, como, por exemplo, os efeitos da poesia e dos cantos religiosos ou o ritmo no mundo infantil e no balbúcio dos bebês; o segundo nível diz respeito ao conhecimento de mundo, que abarca tanto o conhecimento do mundo físico, da interação social humana como os objetivos e as intenções na comunicação. Com relação ao conhecimento de mundo, Luger (2013) destaca que esse nível é fundamental para compreendermos o significado completo de um texto ou conversa. Considerando o foco deste trabalho, buscamos recursos teóricos e modelos linguísticos na subárea da semântica para a tarefa de harmonização a ser desenvolvida.

A história do Processamento de Linguagem Natural é marcada pelos avanços tecnológicos e pela transformação digital. De acordo com Hausser (2014), em 1940, a Ciência da Computação se dividia em computação numeral e não-numeral, em que a

primeira ligada à especialização em cálculos, como na Física, Química e Astronomia, e a computação não-numeral está relacionada aos fenômenos de percepção e cognição. Com essa separação, a computação não-numeral ficou em segundo plano, sendo ocultada pelos estudos desenvolvidos pela computação numeral. Contudo, a computação não-numeral é retomada com o avanço da Inteligência Artificial e das Ciências Cognitivas. Chen (2021) salienta que, nos anos 40, propostas e investigações ligadas à tradução automática e a sistemas de diálogo homem-máquina deram início às tecnologias de PLN. Justamente pela necessidade de se traduzirem línguas para permitir a comunicação entre diferentes povos, inaugurou-se, em 1946, o primeiro computador eletrônico digital, que possibilitava traduzir e processar a linguagem. Nesse período, além do desenvolvimento prático em PLN, os estudos produzidos nas áreas da Linguística, Matemática e Computação contribuíram para a fundamentação teórica no campo de Processamento de Linguagem Natural.

O final da década de 1950 é marcada, ainda, pela crença de que mente e corpo são entidades separadas. No que diz respeito às contribuições linguísticas, em 1957, Noam Chomsky publica seu primeiro livro, “Estruturas Sintáticas”, o qual marca o início do Gerativismo. Conforme discutimos brevemente na subseção anterior, questionando os pressupostos behavioristas, que assumiam que a linguagem era “um fenômeno externo ao indivíduo, um sistema de hábitos gerado como resposta a estímulos e fixado pela repetição” (Kenedy, 2012, p. 128), Chomsky, seguindo a corrente racionalista da linguagem, defende que “a capacidade humana de falar e entender uma língua (pelo menos), isto é, o comportamento linguístico dos indivíduos, deve ser compreendida como o resultado de um dispositivo inato, uma capacidade genética e, portanto, interna ao organismo humano” (Kenedy, 2012, p. 129). Nesse sentido, a linguagem não é mais determinada pelo mundo exterior e pelos estímulos aos quais o homem reage, mas está fundamentada na visão biológica do cérebro e da mente, em que o homem nasce com a linguagem, ou seja, ela é inata.

O inatismo, nessa abordagem, entende a linguagem como algo biológico, em que o ser humano nasce com uma gramática universal inata. Essa concepção evidencia que o homem nasce com uma competência linguística própria para aprender línguas, também chamada de Faculdade da Linguagem. À luz dos pressupostos gerativistas, a gramática universal é composta por regras finitas, e o indivíduo deve ser capaz de criar sentenças infinitas, sendo, assim, possível mostrar que a criatividade é ilimitada. Dessa forma, assumia-se que a linguagem não era previsível, mas sim criativa, já que poderíamos

combinar diversas unidades e gerar novas frases. Contudo, é importante ressaltar que, por esse viés, não se consideram as expressões idiomáticas, por exemplo, uma vez que o princípio da composicionalidade não se aplica a essa vertente (Correa, 2006; Chiavegatto, 2009; Koch; Cunha-Lima, 2009).

Outro ponto discutido por Chomsky se refere à visão modular da cognição e da Faculdade da Linguagem. Baseando-se nos modelos computacionais da época e nos estudos prévios acerca da cognição, Chomsky fundamentava suas análises linguísticas a partir da visão modular, ou seja, acreditava-se que a cognição era separada em módulos, como, por exemplo, o módulo sintático, o semântico e o fonológico, os quais realizavam tarefas de forma autônoma e um desses módulos estava na Faculdade da Linguagem (Kenedy, 2012). Tendo em vista a formalização da linguagem nessa época, a teoria gerativista considera apenas a parte formal da língua, a sintaxe, para suas análises, desconsiderando os demais níveis (Correa, 2006; Chiavegatto, 2009; Koch; Cunha-Lima, 2009).

Com o domínio das correntes racionalistas e empiristas no âmbito de Processamento de Linguagem Natural, entre os anos de 1960 e 1980, a tendência na Ciência da Computação se voltava para pesquisas baseadas no racionalismo, também chamado de PLN simbólico. Segundo Caseli, Nunes e Pagano (2023), o paradigma simbólico buscava expressar a língua a partir de formalismos, como léxicos, regras e linguagem lógica. De acordo com Chen (2021), entre 1968 e 1970, o campo da Inteligência Artificial pretendia transformar o conhecimento humano em uma base de conhecimento formal por meio de sistemas inteligentes baseados em conhecimento. Os sistemas inteligentes fomentaram pesquisas na área de representação do conhecimento, utilizando símbolos pré-definidos em dados estruturados.

Mais adiante, com a revolução cognitiva advinda da Psicologia, os aspectos cognitivos passaram a ser considerados na representação do conhecimento e, nessa conjuntura, Newell e Simon, em 1972, desenvolvem o modelo explicativo do processo cognitivo como atividade humana de resolução de problemas (GPS). Além disso, em 1975, Minsky apresenta a sua teoria de *Frames*, que servia para representar categorias e descrever coisas no mundo (Chen, 2021).

Com os avanços tecnológicos e a crescente popularidade da Internet, em 1990, viu-se o crescimento de dados e informações. Nesse período, conhecido como Paradigma Estatístico, as máquinas passaram a ganhar mais capacidade de memória e processamento, assim como muitos algoritmos de aprendizagem de máquina foram

desenvolvidos (Caseli; Nunes; Pagano, 2023; Deng; Liu, 2018). Com a intensificação do empirismo no campo de Processamento de Linguagem Natural, Caseli, Nunes e Pagano (2023) destacam que o *corpus* é utilizado para a aprendizagem de máquinas e, desse modo, a língua passa a ser representada a partir de modelos probabilísticos, que consideram a frequência de ocorrência. À vista disso, os estudos em PLN começam a usar modelos estatísticos de *data-driven* para tarefas de aplicação, como, por exemplo, para *NER* (Reconhecimento de Entidade Nomeada), *POS tagging* (*Part-of-speech tagging*) e segmentação de palavras (Chen, 2021).

O início dos anos 90 é marcado, na área computacional, pela proposta de Tradução Automática Estatística (*Statistical Machine Translation – SMT*), criada por Peter Brown e pela IBM, que se baseava em um modelo *source-channel*<sup>30</sup> (Chen, 2021; Luger, 2021). Além disso, também surge o método de avaliação BLEU (*Bilingual Evaluation Understudy*), uma forma de medir a diferença na tradução automática usando a tradução feita por humano como referência. Em 2001, Tim Berners-Lee, inventor do WWW, cunhou o termo Web Semântica para tratar da estrutura de dados contidos em sites, identificando os significados semânticos dos conteúdos e das páginas na Internet. Para isso, ele desenvolveu as linguagens semânticas, como a RDF e OWL, para tornar as informações legíveis por máquina. Em 2006, Berners-Lee propôs os princípios de *Linked Open Data* para a publicação de *datasets* na Internet, construídos com tecnologias da Web Semântica (Chen, 2021). No mesmo ano, a empresa Google lançou o primeiro serviço de tradução automática na Internet com um sistema treinando em SMT (*Statistical Machine Translation*), o Google Tradutor.

A transformação digital exigiu máquinas com maior capacidade de memória e processamento para manipular grandes quantidades de dados e criar estruturas mais complexas, como as redes neurais de aprendizagem profunda (*Deep Learning*) (Caseli; Nunes; Pagano, 2023; Deng; Liu, 2018). Chen (2021) aponta que, desde 2010, os modelos de Redes Neurais Profundas (*Deep Neural Networks - DNN*) foram substituindo os modelos de *Machine Learning* e apresentando novas descobertas no âmbito do PLN. De acordo com Caseli, Nunes e Pagano (2023), atualmente, vivemos o

---

<sup>30</sup> Um modelo *source-channel* é um modelo conceitual, que se refere à comunicação linguística feita em duas etapas. A primeira corresponde à geração da mensagem de origem (*source*) e a sua transmissão por meio de um canal com ruídos (*channel*). Nesse processo, inicialmente, modela-se a geração da mensagem original, considerando as probabilidades de sequência de palavras. Na sequência, foca-se na transmissão da mensagem original por meio de um canal, que pode adicionar ruídos ou erros na mensagem-fonte. Esse modelo é muito utilizado na tradução automática justamente para compreender como tais ruídos afetam a tradução (Chen, 2021).

Paradigma Neural, em que as redes neurais utilizam grandes quantidades de dados para aprender um modelo. Todavia, tal aprendizado usa diversas camadas de unidades de processamento para reconhecer os padrões recorrentes. Enquanto no *Machine Learning* os algoritmos indicavam como o aprendizado deveria ser feito, no *Deep Learning*, em virtude da complexidade da arquitetura utilizada, composta por diversas camadas de processamento, não é possível saber com precisão a base que o modelo utilizou para sua aprendizagem (Deng; Liu, 2018). Outra característica do Paradigma Neural diz respeito ao conhecimento da língua, que deixa de ser baseado em símbolos ou regras e passa a ser dado por valores numéricos. Logo “o conhecimento linguístico ou parte do código que tenha produzido um determinado comportamento são praticamente irrecuperáveis, tornando o código opaco, e seu efeito, não previsível (não determinístico)” (Caseli; Nunes; Pagano, 2023, p. 6).

Se, nas abordagens simbólicas, a semântica era representada por uma combinação de símbolos explícitos, no *Deep Learning* passa-se a adicionar as características semânticas ou sintáticas em representação distribuída, como vetores densos, de baixa dimensão e contínuos, chamados de *word embedding* (Chen, 2021). É importante salientar que o conceito de vetor, nesse campo, condiz com um nome matemático para matrizes unidimensionais de número. Anteriormente, tratamos brevemente sobre os *embeddings* e, agora, aprofundamo-nos nesse conceito, visto que é fundamental na área. Como destaca Hagiwara (2021), *word embedding* é um dos conceitos modernos mais importantes no Processamento de Linguagem Natural. Isso ocorre, pois o *embedding*, vetor contínuo, é capaz de representar algo discreto na linguagem. Assim, *word embedding* é uma forma de representar cada palavra com um conjunto de elementos (dependendo do tamanho do conjunto) preenchido com números diferentes de zero. Pensando nisso, pode-se associar o *word embedding* a um tipo de escala numérica em que as palavras são representadas como pontos, de modo que as palavras semanticamente mais próximas estão geometricamente mais próximas.

A título de exemplo, Hagiwara (2021) apresenta as palavras “*dog*” (cachorro), “*cat*” (gato) e “*pizza*” para ilustrar o conceito de *word embedding*. Semanticamente, ‘cachorro’ e ‘gato’ são palavras próximas, visto que ambas designam animais. Contudo, ‘pizza’ não é uma palavra semanticamente próxima de ‘gato’ e ‘cachorro’, uma vez que ela designa um alimento. Tendo isso em mente, cada palavra desse conjunto seria preenchida com um número (diferente de zero) para verificar a aproximação semântica entre elas. É importante destacar que esses índices numéricos são discretos e arbitrários.



Em um primeiro momento, após representar as palavras de forma numérica, teríamos a representação da Figura 27.

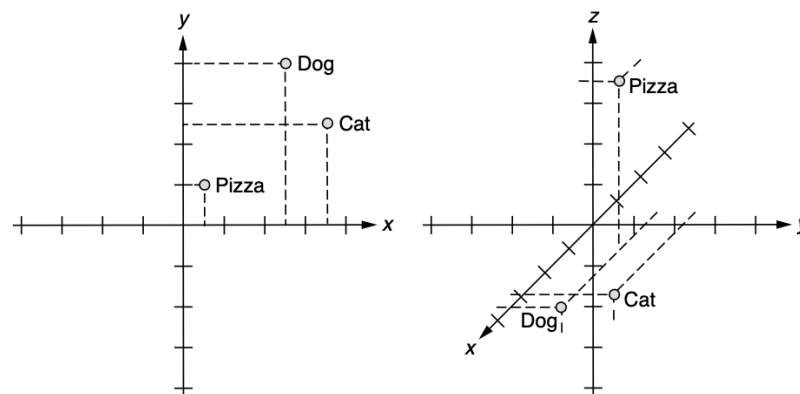
Figura 27 - *Word Embedding* em espaço unidimensional



Fonte: Hagiwara (2021, p. 35).

Nessa versão unidimensional (Figura 27), poderíamos pensar que “pizza” está muito próxima de “dog” e “cat”. Dessa forma, podemos adicionar mais uma dimensão à representação, considerando, inclusive, que computadores lidam melhor com espaços multidimensionais, e, assim, visualizarmos as palavras no formato bidimensional e tridimensional, conforme mostra a Figura 28.

Figura 28 - *Word Embedding* em espaço bidimensional e tridimensional



Fonte: Hagiwara (2021, p. 35).

Na Figura 28, o eixo X representa o conceito de animal, e os eixos Y/Z representam o conceito de alimento. A partir da representação multidimensional das palavras “dog”, “cat” e “pizza”, é possível visualizar o espaço e a proximidade semântica entre as palavras. Posteriormente, é possível, também, adicionar novas palavras e identificar se o valor de cada vetor de palavra é alto o suficiente para dado conjunto de palavras (Hagiwara, 2021).

Tendo em vista o crescente avanço das Redes Neurais Profundas (*Deep Neural Networks* - DNN), que gradativamente substituíram os modelos tradicionais de *Machine*

*Learning*, apresentando grandes contribuições em diferentes disciplinas, com destaque às pesquisas em PLN, o *Deep Learning* alavancou os estudos no campo da Inteligência Artificial (IA). Acompanhando o progresso da IA e seus diferentes paradigmas, viu-se que apenas uma abordagem não seria o suficiente para lidar com as tarefas computacionais e, para tanto, os Paradigmas Híbridos ganharam espaço. Como ressaltam Caseli, Nunes e Pagano (2023), passou-se a combinar a abordagem simbólica com as demais abordagens, evidenciando melhor o conhecimento e, portanto, uma explicação dos passos seguidos pelo algoritmo.

A ascensão e o pleno desenvolvimento de pesquisas e abordagens de Inteligência Artificial e Processamento de Linguagem Natural, sobretudo as técnicas e os métodos de modelagem de linguagem e a possibilidade de acessar grandes quantidades de dados, levaram à elaboração de modelos de linguagem cada vez mais robustos e à criação de aplicações capazes de gerar novos conteúdos. Nesse contexto de inovação, surgem os *Large Models* (LM) (Grandes Modelos – GM) e os *Large Language Models* (LLM) (Grandes Modelos de Linguagem – GML), dando espaço para a Inteligência Artificial Generativa (*Generative Artificial Intelligence* - GenAI).

#### 4.1.4 Modelos de Linguagem e a Inteligência Artificial Generativa

Hoje, muito se fala em modelos de linguagem (LM) e grandes modelos de linguagem (LLM), sobretudo pela disseminação do ChatGPT<sup>31</sup>, LLaMa<sup>32</sup> e Google Bard<sup>33</sup>. Como dito anteriormente, a área da Inteligência Artificial busca automatizar o comportamento inteligente, isto é, desenvolver sistemas inteligentes capazes de processar, raciocinar e agir de forma inteligente. A IA abarca diferentes abordagens, desde a Inteligência Artificial Simbólica, sistemas que buscam representar o conhecimento por meio de regras, sistemas especialistas e programação determinística, até abordagens que se ocupam da criação de sistemas que processam e funcionam de modo semelhantes às sinapses neurais biológicas.

Partindo do princípio de que apenas automatizar o processamento e a inteligência não é o suficiente, observa-se que também é preciso fornecer meios para que a máquina possa aprender. Assim, um subcampo da IA é o *Machine Learning*, como já tratamos na seção 4.1.2, em que o computador, por meio de algoritmos,

---

<sup>31</sup> O chatbot ChatGPT está disponível em: <https://chat.openai.com/>. Acesso em 20 maio 2024

<sup>32</sup> O modelo de linguagem LLaMa está disponível em: <https://ai.meta.com/llama/>. Acesso em 20 maio 2024

<sup>33</sup> O chatbot Google Bard está disponível em: <https://bard.google.com/chat>. Acesso em 20 maio 2024

aprende a identificar padrões em dados brutos e a analisá-los de forma preditiva. Segundo Lima, Pinheiro e Santos (2014), os algoritmos de *Machine Learning* objetivam descobrir relações entre as variáveis de um sistema (com entrada e saída), considerando dados amostrados. Geralmente o desenvolvimento de algoritmos provém de aplicações da Estatística, Física, Engenharia e Computação. Para a aprendizagem de máquina, de modo geral, encontramos três formas: supervisionada, não-supervisionada ou por reforço. Na aprendizagem supervisionada (*Supervised learning*), o modelo é treinado com exemplos rotulados. Para isso, o humano insere os dados, indicando o que é “certo” e o que é “errado” para o modelo. Na aprendizagem não-supervisionada (*Unsupervised learning*), o modelo é treinado sem dados rotulados, isto é, não há uma indicação dos dados, e o modelo busca encontrar padrões e estruturas comuns dos dados. Na terceira forma, a aprendizagem por reforço (*Reinforcement learning*), o modelo é treinado a partir da observação do ambiente e, por meio de tentativa e erro, busca tomar a melhor ação, sendo recompensado ou penalizado de acordo com o resultado apresentado (*Data Science Academy*, 2022).

As técnicas de *Machine Learning* são apropriadas para treinar um modelo com um volume expressivo de informações. Com a vultosa quantidade de dados, as máquinas não davam conta de processá-los e, com o avanço tecnológico e o poder computacional em alta, emergiram novas técnicas computacionais que permitiram novas formas de processar dados. Nesse contexto, surge o *Deep Learning*, subárea do *Machine Learning*, que utiliza “[...] algoritmos para processar dados e imitar o processamento feito pelo cérebro humano.” (Deng; Liu, 2018; *Data Science Academy*, 2022, local. cap. 3). No *Deep Learning*, produzem-se arquiteturas de redes neurais para processar e compreender dados de qualquer natureza, desde a fala humana até objetos visuais. Anteriormente, quando discorremos sobre modelos de linguagem, falamos sobre Redes Neurais Recorrentes (RNN) e *Transformers*.

As Redes Neurais Recorrentes funcionam em ciclos direcionados, permitindo que se volte a acessar os dados de onde se começou. Conforme afirma o *Data Science Academy* (2022, local. cap. 9), as redes neurais recorrentes “são uma maneira muito natural de modelar dados sequenciais. Eles são equivalentes a redes muito profundas com uma camada oculta por fatia de tempo; exceto que eles têm os mesmos pesos em cada fatia de tempo e recebem entrada em cada fatia”. Com isso, um modelo baseado em Redes Neurais Recorrentes aceita uma entrada de tamanhos variáveis e segue a ordem dos componentes da entrada, a fim de induzir um vetor que represente uma

sequência (Paes; Vianna; Rodrigues, 2023). Dessa forma, tal modelo irá identificar uma palavra a partir da sequência de palavras anteriores. De modo geral, um modelo de linguagem com RNN segue as seguintes etapas: (i) calcula o vetor de *embedding*; (ii) calcula a saída da camada escondida; (iii) calcula a saída; (iv) calcula a distribuição de probabilidade; (v) resgata a palavra com o maior valor de probabilidade; e (vi) o processo continua até o modelo encontrar um *token*<sup>34</sup> de fim de sequência ou até alcançar uma saída máxima (Paes; Vianna; Rodrigues, 2023; Data Science Academy, 2022).

Os *Transformers*, por sua vez, são arquiteturas de redes neurais, desenvolvidos em 2017, e apresentam características distintas, favorecendo a crescente preferência por eles na computação. Um dos diferenciais de um *Transformer* é a possibilidade de realizar tarefas *sequence-to-sequence* (*seq2seq*), isto é, ele transforma uma determinada sequência de elementos em outra sequência, obedecendo a princípios sintáticos e semânticos. Nesse sentido, as sentenças dependem da sequência, uma vez que a ordem (de letras, palavras e textos) é crucial para a compreensão de uma sentença (Kublik; Saboo, 2022; Ozdemir, 2023). Para isso, geralmente, encontramos dois componentes: um codificador (*encoder*), que processa a sequência de entrada e codifica como um vetor (vetor de contexto), e um decodificador (*decoder*), que recebe e processa o vetor e transforma na sequência de saída. A título de exemplo, o Google Tradutor utiliza o modelo *seq2seq*, pois esse tipo de arquitetura é ótimo para tradução automática, já que exige uma entrada (uma sequência) e uma saída (outra sequência) (Paes; Vianna; Rodrigues, 2023).

Outro ponto fundamental do *Transformer* é o mecanismo de atenção. Tendo em vista que nem sempre é possível captar todas as nuances para resolver uma tarefa, sobretudo se a sequência de entrada for muito longa, utiliza-se a estratégia de atenção. O mecanismo de atenção é uma técnica que busca imitar a atenção cognitiva, olhando para a sequência de entrada, parte por parte, e, com base em probabilidade (pesos de atenção), decide em cada passo quais outras partes da sequência são importantes (Kublik; Saboo, 2022; Paes; Vianna; Rodrigues, 2023). Nesse tipo de arquitetura, utiliza-se um mecanismo de atenção adicional, a auto-atenção (*self-attention*). Ela permite encontrar relações entre diferentes palavras da entrada a partir da estrutura

---

<sup>34</sup> *Token* se refere à menor unidade de significado criado pela divisão de uma frase ou trecho de texto em unidades menores, sendo ele a entrada básica para um LLM. Uma sequência de *tokens* consecutivos é chamada de *n-gram* (Ozdemir, 2023).

sintática e contextual da sentença, considerando os *embeddings* e seus pesos. Dessa forma, é possível que o *Transformer* foque no que é relevante na sequência, como conectar duas palavras em uma relação semântica entre elas, quando as palavras em si não carregam nenhuma marcação apontando para essa relação (*Data Science Academy*, 2022; Kublik; Saboo, 2022; Paes; Vianna; Rodrigues, 2023; Ozdemir, 2023).

De acordo com Paes, Vianna e Rodrigues (2023), os *Transformers* apresentam um nível de paralelismo no processamento da sequência de entrada. Isso ocorre devido à complexidade de uma sentença, que conta com contexto e palavras com diferentes significados e referências. No entanto, apenas o mecanismo de atenção não é suficiente para captar essas características. Assim, eles incluem uma camada de atenção em múltiplas versões (*multi-heads attention*). Essa camada de atenção trabalha com um tipo diferente de *embedding*, que lida com a ordem e a posição das palavras no codificador. No decodificador, a camada de auto-atenção, chamada de mascarada, que não tem acesso aos demais *embeddings*, comunica-se com o codificador. Dessa forma, a atenção codificador-decodificador combina as conexões entre palavras da frase-fonte (frase de entrada) e as palavras da frase-alvo (frase de saída).

Mas por que falamos de *Transformers*? Os *Large Language Models* (LLMs) são modelos de Inteligência Artificial que, geralmente, são derivados da arquitetura *Transformer*. Eles podem apresentar apenas o codificador, como o modelo BERT (*Bidirectional Encoder Representations for Transformers*), apenas o decodificador, como o modelo GPT (*Generative Pre-Trained Transformer*), ou a arquitetura completa, como o modelo T5, dependendo do tipo de tarefas que pretende realizar, conforme podemos observar no Quadro 3.

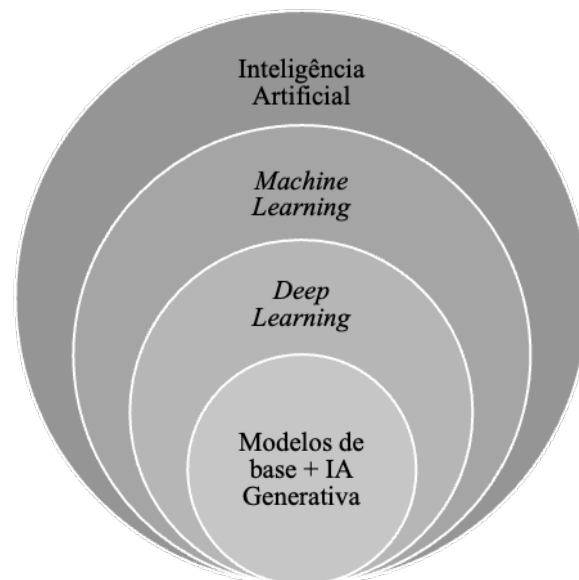
Quadro 3 - Modelos de *Transformers*

Modelo	Descrição
Modelo BERT	Nesse tipo de modelo, mascaram-se alguns <i>tokens</i> de entrada e se busca prever o próximo <i>token</i> a partir do restante dos <i>tokens</i> , capturando as relações contextuais entre eles. É eficaz para classificação de texto.
Modelo GPT	Prevê o próximo <i>token</i> com base nos <i>tokens</i> anteriores. É eficaz na geração de textos livres e coerentes, seguindo um determinado contexto.
Modelo T5	Utiliza o codificador e o decodificador para ser mais flexível e versátil na geração de texto. Assim, gera textos mais criativos e em diferentes contextos, uma vez que o codificador auxilia na captura adicional de contexto.

Fonte: Adaptado de Ozdemir (2023, local. cap. 1).

Tais modelos são treinados com grandes quantidades de dados, possibilitando capturar toda a complexidade e as nuances da linguagem natural, e são chamados modelos de base (*foundation models*). Os modelos de base, como os descritos acima, são redes neurais treinadas com grandes volumes de dados, medidos em bilhões de parâmetros e treinados com bilhões de dados generalizados e não rotulados (Ozdemir, 2023). Desse modo, a aprendizagem do modelo é auto-supervisionada, ou seja, os exemplos de rótulos são fornecidos na própria estrutura dos dados, permitindo que o modelo descubra os padrões escondidos ou a estrutura dos dados, prevendo a próxima palavra. À vista disso, um modelo de base é capaz de realizar diversas tarefas gerais, desde a compreensão da linguagem até a geração de textos, imagens e conversas em linguagem natural, pois apresenta distintas capacidades, como classificação de texto, extração de entidades, sumarização, geração de código, mecanismo de busca, pergunta e resposta e criação de novos conteúdos. É justamente por ser capaz de gerar novos conteúdos e apresentar tantas capacidades que chamamos esse tipo de inteligência como Generativa (*Generative Artificial Intelligence*). Para melhor compreensão de todas as áreas e subáreas descritas acima, podemos visualizar o campo da Inteligência Artificial e suas subáreas, como ilustra a Figura 29.

Figura 29 - Área e subáreas da Inteligência Artificial



Fonte: Elaborada pela autora.

Abrangendo todos os campos, encontramos a Inteligência Artificial, que busca automatizar e processar o comportamento inteligente. Em seguida, como subárea da IA, temos o *Machine Learning*, que desenvolve técnicas e abordagens focadas na

aprendizagem de modelos. Mais a fundo, o *Deep Learning*, subcampo do *Machine Learning*, que objetiva elaborar técnicas computacionais especializadas que imitam os processos do cérebro humano, criando redes neurais; e, finalmente, encontramos os modelos de base e a Inteligência Artificial Generativa, que, por meio dos modelos criados com redes neurais, os *Transformers*, criam novos conteúdos e exibem diversas capacidades, como extrair entidades, classificar de texto, gerar código, dentre outras possibilidades.

Como vimos, os modelos de base são derivados da arquitetura *Transformer* e treinados com massivas quantidades de dados para desempenhar tarefas gerais e, posteriormente, podem ser treinados com conjuntos de dados menores para tarefas mais específicas. Uma classe de modelos de base são os *Large Language Models* (LLMs). Um dos LLMs mais conhecidos é o GPT, criado pela OpenAI<sup>35</sup> em maio de 2020 e liberado para acesso em junho de 2020. Os modelos generativos são capazes de realizar tarefas gerais baseadas na linguagem, como, por exemplo, classificar e gerar textos de diferentes estilos e com diversas finalidades, com alta precisão e fluência (Kublik; Saboo, 2022; Ozdemir, 2023).

Por trás do LLM, há um trabalho de modelagem de linguagem, subárea de Processamento de Linguagem Natural, que envolve a criação de modelos estatísticos em *Deep Learning* para prever a probabilidade de *tokens* em um vocabulário específico, isto é, um conjunto limitado e conhecido de *tokens*. No que se refere à modelagem de linguagem, Ozdemir (2023) salienta que podemos encontrar dois tipos de tarefas:

- (i) tarefas de codificação automática, em que os modelos são treinados para reconstruir a frase original a partir de uma versão corrompida da entrada – tarefa que corresponde ao codificador (*encoder*) do *Transformer* e tem acesso à entrada completa sem as máscaras –, criando uma representação bidirecional de toda a frase, e sua função principal é classificar frases ou *tokens* (podendo, também, ser ajustados para gerar textos);
- (ii) tarefas de autorregressão, em que os modelos são treinados para prever o próximo *token* em uma frase, baseando-se nos *tokens* anteriores – corresponde ao decodificador (*decoder*) do *Transformer* e uma máscara aplicada à frase completa para que as entradas de atenção possam ver apenas

---

<sup>35</sup> A OpenAI é “uma empresa de pesquisa e implantação de IA. Nossa missão é garantir que a inteligência artificial geral beneficie toda a humanidade”. SOBRE. In.: OPENAI. Califórnia, 2024. Disponível em: <https://openai.com/>. Acesso em: 14 jan. 2024.

os *tokens* prévios –, e são ideais para gerar textos, como, por exemplo, o GPT (Ozdemir, 2023).

Ao analisarmos o histórico da criação e o aprimoramento dos modelos GPT, podemos perceber que as diferentes capacidades do modelo e a quantidade de parâmetros e dados são essenciais para que o ChatGPT, por exemplo, possa performar e ser excelente em geração de textos. De acordo com Ozdemir (2023), o modelo GPT-1, criado em 2018, provou que modelos de linguagem podem ser pré-treinados de forma eficiente, contribuindo para a generalização e podendo compreender a linguagem. Para esse modelo, utilizou-se um *dataset* composto pelo *Book Corpus*, constituído por 117 milhões de parâmetros. Posteriormente, em 2019, a OpenAI lançou o GPT-2, evidenciando que o modelo melhora sua capacidade de realizar tarefas a partir do treinamento com um conjunto de dados ainda maior e com mais parâmetros. Desse modo, o GPT-2 foi treinado com o *dataset* da *WebText* e com 1.5 bilhão de parâmetros, dez vezes mais do que o conjunto de dados utilizado no modelo GPT-1. Se compararmos os dois modelos desenvolvidos até então, o GPT-2 tem como diferencial a capacidade de realizar tarefas múltiplas, sem receber nenhum exemplo de treinamento para isso. Em 2020, lançou-se o GPT-3, com 175 bilhões de parâmetros e treinando com um conjunto de cinco *corpora* diferentes. Considerando isso, ele pode executar tarefas para as quais não foi diretamente treinado, como calcular e escrever consultas SQL (*Structured Query Language* – Linguagem de Consulta Estruturada), bem como possuir recursos para escrever artigos que são indistinguíveis dos produzidos por humanos (Kublik; Saboo, 2023). O último lançamento da OpenAI foi o GPT-4, em 2023, com 170 trilhões de parâmetros e é uma aplicação multimodal, permitindo receber como entrada tanto textos como imagens.

A interação entre o modelo GPT e o usuário é feita por meio de *prompts*, ou seja, o ser humano acessa o ChatGPT e fornece instruções explícitas, guiando o modelo a produzir respostas desejadas. Nesse sentido, a instrução pode ser feita de diferentes formas e apresentar diferentes informações. Contudo, vale destacar que, dependendo da disposição da informação, as respostas podem variar. Liu, Zeng, Meng e Zhou (2023) apontam que a sequência da instrução no *prompt* de uma tarefa é crucial para obter uma resposta melhor e mais precisa. Os autores salientam que alguns aspectos requerem mais atenção, como o tamanho da instrução e a posição das informações. Após os experimentos realizados em LLMs, Liu *et al.* (2023) constataram que a instrução no *prompt* deve considerar, primeiro, a sentença de *input* e, depois, a instrução da tarefa



específica. Outro ponto sobre a instrução se refere à inserção de exemplos, contextos, dicionários e outras informações que possam auxiliar o modelo a fornecer respostas mais acuradas. Desse modo, ao apresentar exemplos próximos do que se deseja, o LLM pode produzir uma resposta mais próxima do que se espera.

Embora os produtos desenvolvidos pela OpenAI sejam uma revolução na IA, alguns pontos ainda precisam de atenção e apresentam limitações, como a situação denominada pelo termo alucinação. Pela sua capacidade de gerar conteúdos e novos textos, os LLMs podem criar respostas plausíveis, mas falsas, isto é, o conteúdo da resposta pode parecer correto, mas algumas informações apresentadas ali não estão de acordo com a realidade. A alucinação é um ponto de alerta, visto que, atualmente, muito se “copia” do modelo, sem verificar a veracidade do conteúdo. Essa característica do LLM depende muito da forma como foi perguntado ao modelo, como vimos na instrução de *prompt*, mas também do treinamento da aplicação, visto que o modelo é treinado em determinado tempo e, se não houver mais treinos, os dados mais atuais não serão considerados.

Uma forma de minimizar o problema de alucinação e fornecer insumo para o modelo é a partir da adição de *Knowledge Graphs* (KGs) e Ontologias<sup>36</sup> no *prompt*. Enquanto os LLMs ainda apresentam limitações com relação a sua habilidade de raciocinar sobre relações complexas entre entidades e conceitos, as ontologias e KGs podem contribuir para essa problemática ao oferecer conhecimento externo que pode ser usado para raciocinar e gerar textos mais precisos e eficientes. Essa união entre LLM e KG/ontologias permite que o processamento de linguagem natural e os modelos possam criar conteúdos mais bem estruturados (Goyal; Bhattacharyya, 2023). Moiseev, Dong, Alfonseca e Jaggi (2022) discutem que a integração de *Knowledge Graphs* em LLMs pode melhorar a sua habilidade de responder perguntas mais complexas, que exigem um conhecimento prévio. Além disso, com as informações provenientes do KG, o modelo de linguagem pode identificar as entidades e relações mais relevantes e aplicá-las para a geração de respostas mais significativas.

Um dos objetivos deste trabalho é construir um *Knowledge Graph* do ERP, considerando os *frames* semânticos na sua estrutura, para compreender a terminologia do *software*, bem como para explorar técnicas e estratégias de *prompt* em LLM para tratar da tarefa de harmonização de termos e conceitos seguindo a abordagem

---

<sup>36</sup> Na subseção 4.2, trataremos da representação do conhecimento, abordando as Ontologias e *Knowledge Graphs*.

semântico-terminológica. Por isso, a seguir, discutiremos a representação do conhecimento, aprofundando-nos nos conceitos de Ontologia, *Knowledge Graphs* e sua relação com os *Frames*.

#### 4.2 Representação do Conhecimento: Ontologias, *Knowledge Graphs* e *Frames*

Historicamente, a representação do conhecimento (RC) estava associada à lógica, como na Matemática, em que a representação do conhecimento se ligava ao raciocínio correto, gerando linguagens formais com regras de inferência (Luger, 2013; Sowa, 2000). No campo da Psicologia e da Linguística, a RC era vista como a caracterização da natureza a partir do conhecimento humano, preocupando-se, principalmente, em descrever e compreender como os seres humanos adquirem, associam e usam o conhecimento de mundo (Weller, 2010; Luger, 2013). Na Ciência computacional, a representação do conhecimento se refere aos estudos e às pesquisas que se voltam para a área de Inteligência Artificial, discutindo meios para estruturar o conhecimento humano e que este seja legível por máquina. Atentando para as diferentes formas de compreender a representação do conhecimento, percebe-se que esta área é interdisciplinar e, por isso, apresenta diferentes definições. De acordo com Weller (2010), a representação do conhecimento pode ser compreendida como uma abordagem que discute como estruturar, armazenar e recuperar a informação de forma precisa e eficaz.

Uma forma inicial de representar o conhecimento foram as redes semânticas. Tais estruturas eram compostas por nós e arestas interconectadas. As redes semânticas são formalizações das teorias associacionistas do conhecimento. Luger (2013) conceitua as teorias associacionistas como ligadas à tradição empirista, as quais definem o significado de um objeto em termos de uma rede de associações com outros objetos. Nesse sentido, os seres humanos percebem um objeto no mundo, e essa percepção é mapeada, inicialmente, como um conceito que faz parte do nosso conhecimento de mundo e está conectado por meio de relacionamentos a outros conceitos. Essa associação entre os conceitos, para o homem, é organizada de forma hierárquica, mantendo a informação em níveis adequados e mais altos da taxonomia.

Na composição das redes semânticas, os nós representavam conceitos de entidades, objetos, eventos, atributos e estados; as arestas eram chamadas de relações

conceptuais com tipos específicos de relações, as quais representavam os relacionamentos entre os nós. Como afirma Luger (2013, p. 193),

Em geral, a compreensão da linguagem natural requer a compreensão do senso comum, os modos como os objetos físicos se comportam, as interações que ocorrem entre os seres humanos e os modos como as instituições humanas estão organizadas. Um programa de linguagem natural precisa compreender intenções, crenças, raciocínio hipotético, planos e objetivos. Devido a esses requisitos, a compreensão de linguagem natural tem sido sempre uma força propulsora para as pesquisas de representação do conhecimento.

Desse modo, a rede semântica possibilita que estructuremos os conceitos e representemo-los linguisticamente. As primeiras implementações de redes semânticas na computação foram na área de tradução automática, em 1960 (Sowa, 1991). No final dos anos 60, Ross Quilian propôs um programa baseado em redes semânticas, o qual definia as palavras de forma semelhante aos dicionários (Luger, 2013). O recurso de Quilian definiu muitas características acerca do formalismo das redes semânticas, uma vez que os arcos e elos rotulados, as heranças hierárquicas e as inferências entre os elos associativos permitiam que a máquina pudesse compreender tal estrutura. A formalização proposta por Quilian apresentava uma limitação quanto às habilidades para tratar das complexidades de muitos domínios, já que muitos elos representavam associações genéricas e não forneciam uma base real para a estruturação dos relacionamentos semânticos. Por esse motivo, busca-se formalizar as relações semânticas básicas da linguagem natural, e, conseqüentemente, os fundamentos do conhecimento alcançam maior generalidade e consistência.

Segundo Luger (2013), ainda no final dos anos 60, percebeu-se a necessidade de se estabelecerem relacionamentos padrão e uma alternativa para isso foi encontrada na estrutura de casos dos verbos em inglês. Em seu artigo publicado em 1967, Charles Fillmore propôs que a noção gramatical “caso” fosse considerada no componente básico da gramática de cada língua. Nos estudos gerativistas, a sintaxe era vista como ponto principal para a compreensão e produção linguística, e o componente semântico era deixado de lado. Contudo, Fillmore (1967) aponta que muitos aspectos semânticos são relevantes, sobretudo na estruturação profunda da língua, em que uma frase é constituída por um verbo e por uma ou mais frases nominais, sendo cada uma delas ligada a um verbo em uma relação de caso (Fillmore, 1967). Para Fillmore, os casos se referem aos papéis semânticos, como agente, localização, instrumento, tempo, objeto,

dentre outros; assim, o verbo seleciona diferentes casos, formando um quadro de casos. Em outras palavras, um quadro de casos corresponde a uma sentença representada como um nó verbal e composta por vários elos de casos que ligam tais nós, os quais representam outros participantes da ação. A estrutura, então, apresenta informações fundamentais sobre a valência semântica dos verbos, adjetivos e substantivos (Fillmore, 1967; Luger, 2013).

Computacionalmente, o quadro de casos permite que o programa, ao analisar uma sentença, encontre o verbo e recupere seu quadro de casos em sua base de conhecimento. Por conseguinte, ele ligará os valores do agente, objeto e demais papéis semânticos aos nós apropriados no quadro de casos. A partir dessa sistemática, é possível captar aspectos mais profundos da linguagem natural, utilizando a relação verbo-sujeito e verbo-objeto (Luger, 2013).

Com a necessidade de se estruturar melhor a representação do conhecimento e como as redes semânticas não tinham poder para lidar com informações mais específicas, Marvin Minsky propôs a teoria de *Frames* (quadros), em 1975. Os *frames*, nessa abordagem, são vistos como esquemas representacionais do conhecimento, que estruturam e organizam as conexões implícitas de um dado domínio de problema de forma explícita. Com isso, é possível organizar o conhecimento em unidades mais complexas, que refletem a organização de objetos de uma área. Em seu artigo publicado em 1974, Minsky argumenta que as teorias ligadas à Inteligência Artificial e à Psicologia da época não consideravam o conhecimento como um todo, e isso afetava o raciocínio, a memória e a percepção na elaboração de sistemas. Contrariando a visão de compreender o conhecimento separadamente, Minsky apresenta o *frame* como uma rede de nós e relações. Assim, um *frame* é

uma estrutura de dado para representar uma situação estereotipada, como estar em um certo tipo de sala de estar, ou ir a uma festa de aniversário de criança. Conectado a cada *frame* estão vários tipos de informação. Algumas dessas informações são sobre como se usa o *frame*. Outras são sobre o que se pode esperar que aconteça em seguida. E outras são sobre o que se pode fazer se essas expectativas não são confirmadas (Minsky, 1974, p. 111-112).

Nessa estrutura de *frames*, encontramos, no topo, os *frames* fixos, os quais representam coisas que são sempre “verdade” sobre uma suposta situação. Nos níveis mais baixos da estrutura, encontramos os “terminais”, repartições (*slots*) compostas por

informações específicas. Cada terminal pode apresentar condições particulares que os atributos exigem, gerando *subframes* (Minsky, 1975; Luger, 2013).

Se, por um lado, as redes semânticas permitiam que organizássemos o conhecimento em nós e arestas, mas não compreendiam toda a complexidade dos conceitos e demandavam da interpretação do usuário para entender a representação de um objeto/conceito, por outro lado, os *frames* de Minsky tornam mais clara a descrição de objetos/conceitos. Isso se verifica quando conectamos *frames* relacionados com suas descrições detalhadas, geramos um sistema de *frames* e, conseqüentemente, podemos analisar diferentes visões de um mesmo objeto/conceito. Desse modo, essas estruturas se tornam mais poderosas do que as redes semânticas, pois possibilitam que objetos complexos sejam representados como um único *frame*, em vez de uma grande estrutura em forma de rede. Um sistema de *frames* é uma alternativa para organizar e estruturar uma base de conhecimentos, mas essa abordagem ainda exige muito da habilidade e intuição do programador para a sua construção.

Como vimos anteriormente, Charles Fillmore se inspirou nos *frames* de Marvin Minsky para desenvolver a teoria da Semântica de *Frames*. No contexto linguístico, afastando-se dos pressupostos estruturalistas e gerativistas, Fillmore afirma que os *frames* estão relacionados às situações definidas por questões culturais e sociais, experiências humanas e estruturas institucionais no nível conceptual. Dessa forma, ele cria um modelo para análise semântica que aborda tanto as questões contextuais como as experienciais para poder compreender as cenas e os *frames*.

Posterior aos *frames* de Minsky, um caminho para a representação do conhecimento foi a criação de ontologias. Para o desenvolvimento de agentes resolvedores de problemas em ambientes complexos, geralmente é preciso implementar diversos esquemas de representação do conhecimento. Assim, cada esquema pode ser considerado uma ontologia (Luger, 2013). Conforme afirmam Malucelli e Bax (2012), a ontologia é uma noção interdisciplinar, que trabalha com conceitos e teorias que auxiliam na construção de conceptualizações especializadas de um determinado domínio.

Sánchez, Cavero e Martínez (2007) salientam que o conceito de ontologia, na Computação, originou-se da Filosofia. Na Grécia Antiga, Aristóteles já discutia sobre a essência das coisas no mundo e como isso gera tudo o que nos rodeia, conferindo a propriedade de “ser” aos objetos na realidade (Schalley, 2019). No século XVII, Christian Wolff, ao tratar da metafísica, distingue a metafísica geral – também chamada

de Ontologia – da metafísica especial. Nesse sentido, a ontologia podia ser compreendida sob duas perspectivas. A primeira se referia à abordagem que buscava perceber os aspectos essenciais que nos permitem nomear as coisas; e a segunda investigava as hierarquias e classificações dos seres, mostrando que os níveis mais altos da hierarquia remetiam às propriedades gerais, e os níveis mais baixos correspondiam às características mais específicas. Tendo isso em vista, na Filosofia, a ontologia estava conectada aos estudos metafísicos sobre a natureza e as relações de existência (Gandon, 2010).

Para se trabalhar com a representação do conhecimento pela perspectiva computacional, Gandon (2010) destaca que a computação recorreu aos questionamentos existenciais da Filosofia para tratar das ontologias e da possibilidade de arquitetar seus conceitos e relações. De acordo com Müller (2015), nas ciências computacionais, o termo ontologia está mais ligado à construção ou à estruturação de um determinado domínio, que seja capaz de descrever o mundo ou uma situação, representando o conhecimento, para fins computacionais. Na visão epistemológica, o conhecimento é feito de conceitos, sendo o conceito a representação do significado de algo, ou seja, a representação mental de um objeto quando o homem pensa sobre tal objeto. Considerando isso, Gruber (1992), do ponto de vista computacional, define a ontologia como uma especificação explícita de uma conceitualização, que apresenta uma visão abstrata e simples de como vemos e compreendemos o mundo.

Sowa (2000) define a ontologia como a especificação de tipos de entidades que existem ou podem existir em algum domínio ou área temática. Pensando nisso, a ontologia, então, pode ser vista como um conjunto de conceitos e tipos de relações organizadas de forma hierárquica (Lacasta; Nogueras-Iso; Zarazapa-Soria, 2010). Em sua definição, Sowa (2020) também discorre sobre a modelagem do domínio e afirma que a linguagem utilizada na construção de uma ontologia de domínio é fundamental para que se consiga estruturá-la. Dessa forma, Sowa (2020), implicitamente, adiciona a língua como fator importante na ontologia (Schalley, 2019).

Com a relevância da língua na estrutura ontológica, torna-se imprescindível discutir sobre a ontologia pelo panorama linguístico. Como explicita Schalley (2019), o uso de ontologias na Linguística cresceu na última década, sobretudo pelas tecnologias da Web Semântica e pelos desenvolvimentos na área de IA e PLN. Atualmente, as ontologias são vistas como estruturas de representação ou modelagem para pesquisa linguística e, em conjunto com os recursos computacionais, possibilitam novos métodos

de análise da língua e do conhecimento. Além disso, com a diversidade linguística e terminológica, as ontologias fornecem meios para organizar toda a estrutura da língua geral e específica, bem como o conhecimento de áreas especializadas.

Pelo viés da Linguística, encontramos a ontologia linguística. Partindo do pressuposto de que uma ontologia tem por objetivo extrair o conhecimento compartilhado de um domínio, identificar termos e relações desse conhecimento e formalizar toda essa estrutura, fornecendo um entendimento consensual do conhecimento (Hang *et al.*, 2010), para a construção de uma ontologia, é necessário refletir sobre a conceptualização – processo que leva em conta a extração e a generalização de uma informação a partir da experiência de alguém e a sua visão de mundo. Conectando isso à ontologia linguística, percebe-se que ela irá, de certo modo, refletir um determinado grau de consenso em experiências linguísticas e culturais (Wu; Yuan, 2019). Na visão de Schalley (2019), a ontologia linguística pode ser descrita como uma rede interconectada de conceitos, que tornam explícitos, classificados e organizados suposições e termos de um domínio a partir da conceptualização humana.

No campo de Processamento de Linguagem Natural, há um contraponto no que diz respeito às ontologias linguísticas e às ontologias formais (ou conceituais) (Vossen, 1998; Palmer, 2001; Farrar; Bateman, 2005). De acordo com os autores, a ontologia é linguística ou formal a depender do viés pelo qual é estruturada, visto que é considerada linguística quando se preocupa em registrar os conceitos lexicalizados de uma língua, tratando-se de um "[...] inventário dos sentidos de uma dada língua" (Di Fellipo, 2008, p. 470). As ontologias formais se concentram em organizar conceitos. Nessas ontologias, o que ocorre é a inserção de níveis artificiais para que seja possível a representação de uma estrutura controlada dos conceitos que pertencem a um domínio de conhecimento. A grande diferença entre ambas as ontologias está no tamanho e no grau de conceptualização, uma vez que as ontologias formais utilizam técnicas lógicas para a descrição do conhecimento, enquanto as linguísticas descrevem a semântica de suas unidades lexicais com base em diversas perspectivas de organização conceitual. (Müller, 2015, p. 25).

Outra abordagem que merece destaque é a Ontolinguística. Schalley e Zaefferer (2007) apresentam o conceito de Ontolinguística para tratar de ontologias como "uma rede de conceituações interconectadas dos fenômenos que constituem o mundo" (Schalley; Zaefferer, 2007, p. 3). A Ontolinguística insere-se no campo da Linguística Cognitiva, assim como a teoria da Semântica de *Frames*, visto que o conhecimento

linguístico, nessa corrente, é abordado como um tipo de conhecimento ontológico, baseando-se no princípio de que existem universais mentais (Schalley; Zaefferer, 2007). Seguindo o conceito de Ontolinguística, encontramos a noção de Ontolex. Do ponto de vista ontológico, a base da construção de uma ontologia são conceitos e relações. Uma preocupação semelhante centrada em termos e relações é encontrada em recursos lexicais computacionais, semelhantes as ontologias linguísticas. Sendo assim, conceitos (ou palavras) e relações são o ponto de partida para se trabalhar com ontologias e recursos lexicais. Esse paralelismo define a interface da Ontolex (Prévo, 2010). A análise e construção ontológica lidam com conceitos fundamentados na representação do conhecimento. Por outro lado, as ontologias linguísticas partem de um vocabulário existente e apresentam uma organização útil para esses termos baseada em semântica. O trabalho situado na interface da Ontolex deve, portanto, encontrar a melhor integração de ambas as abordagens, tratando-se de uma combinação equilibrada entre elas. Essa interface também se mostra importante no *design* de recursos multilíngues, já que essas ferramentas são tipicamente constituídas a partir de vários recursos monolíngues mapeados para uma interlíngua (Hang *et al.*, 2010).

Essa conceituação das ontologias e da representação de conhecimento na Linguística pode ser ainda muito mais aprofundada e detalhada quando levamos em consideração outras teorias de cunho mais puramente linguístico, como é o caso de teorias semânticas. Tais teorias enriquecem o escopo das ontologias, oferecendo uma camada contextual e situacional que, muitas vezes, apenas a semântica formal, por exemplo, não proporciona. A título de exemplo, discorreremos sobre a Semântica de *Frames*, teoria de Charles Fillmore, que se inspirou nos *frames* de Marvin Minsky, a fim de propor uma modelo para análise semântica. Pensando nisso, uma forma de introduzir essa camada é a partir do uso de *frames* semânticos.

Conforme afirma Müller (2015), ainda que a Linguística não se dedique tão profundamente às questões da ontologia, podemos observar que há uma preocupação em buscar soluções para tratar do aspecto semântico com o objetivo de estabelecer as relações entre os conceitos. Para Müller (2015), uma forma de abordar esse ponto é utilizando todo o potencial dos *frames* semânticos para se trabalhar na descrição de tal aspecto. Do ponto de vista computacional, seguindo as definições e a abordagem de Minsky, uma ontologia baseada em *frames* é caracterizada como o *frame* sendo representado como um conceito (ou classe) e este é definido pelo seu identificador e por um número de elementos (*slots*) e cada elemento corresponde a um atributo que os



membros da classe apresentam. Nesse contexto, essa estrutura irá incluir classes e propriedades que são especificadas nos níveis mais gerais e são herdadas por subclasses e instâncias (Lacasta; Nogueras-Iso; Zarazapa-Soria, 2010). Para ilustrar essa relação *frames-ontologia*, as ferramentas Protégé e a Ontolingua *Servers* são recursos utilizados para a construção de ontologias que se baseiam em *frames*.

Mais além, na visão de Scheffczyk, Baker e Narayanan (2010), combinar recursos lexicais baseados no conhecimento de mundo com as ontologias é uma etapa crucial para aprimorar o raciocínio pela linguagem natural, sobretudo na Web Semântica. Para exemplificar essa relação, os autores utilizam a FrameNet, um recurso computacional que se inspirou na abordagem fillmoreana para a sua elaboração. A FrameNet<sup>37</sup> é um recurso lexical com alto detalhamento semântico para realizar experimentos computacionais no campo da representação do conhecimento. Tal recurso objetiva criar uma base de dados lexicais capaz de ser lida tanto por seres humanos como máquinas. Esse produto é uma ferramenta utilizada por pesquisadores na área de PLN, visto que fornece informações semânticas de sentenças anotadas, que servem para extração de informações, tradução automática, análise de sentimentos e reconhecimento de eventos. Scheffczyk, Baker e Narayanan (2010) refletem sobre as aproximações entre a FrameNet e a ontologia, uma vez que os *frames* permitem que os sistemas façam inferências em textos de linguagem natural por meio das relações semânticas. Embora os tipos semânticos da FrameNet sejam muito superficiais, não apresentando muitas relações entre eles, o recurso lexical possibilita que se faça um treinamento de dados para a aprendizagem de máquina.

Com a aproximação entre *frames* e ontologias, Scheffczyk, Baker e Narayanan (2010) apresentam uma possível metodologia para lidar com esse tipo de representação. Para eles, primeiramente, é necessário representar os *frames*, tipos e tipos semânticos como classes. O segundo passo é modelar as relações entre os *frames* e os elementos de *frames* por meio de restrição de propriedade existencial nas classes – as restrições de classe, nessa visão, são herdadas. Por fim, a terceira etapa consiste em representar os *frames* e preencher os elementos de *frames* em textos anotados como instâncias do *frame* adequado e das classes de preenchimento de elementos de *frame*.

Em sua tese de doutorado, Müller (2015) propôs uma metodologia baseada na Semântica de *Frames* para a construção de ontologias no contexto jurídico. Em sua

---

<sup>37</sup> Os recursos da FrameNet, bem como os materiais teóricos utilizados no projeto estão disponíveis em: <https://framenet.icsi.berkeley.edu/> Acesso em: 14 jan. 2024.

proposta, a autora mostra que, ao adicionar as informações provenientes dos *frames*, a ontologia apresenta uma descrição mais detalhada dos conceitos do domínio. A metodologia de Müller (2015) consiste em cinco etapas. De início, é necessário planejar e especificar os requisitos, definindo o domínio e o escopo da ontologia. Em seguida, na etapa de aquisição do conhecimento, compila-se um *corpus* de pesquisa para a seleção de conceitos e para a estruturação da ontologia. A terceira etapa consiste na conceitualização. De acordo com Müller (2015), essa etapa é fundamental para uma ontologia com *frames*, pois é aqui que as classes, subclasses e relações são descritas. Nesse momento, utilizam-se três abordagens: a abordagem *Top-down*, em que se parte de um mapa conceitual para a construção dos *frames*; a abordagem *Middle-out* para estruturar as classes e subclasses (taxonomia) e as relações não taxonômicas; e a abordagem *Bottom-up* para, por meio de uma lista de palavras extraída do *corpus* compilado, adicionar termos do domínio que irão compor toda a estrutura e fornecer exemplos em uso para ilustrar os conceitos.

Ainda no âmbito das ontologias, mas focando nas questões terminológicas, Lacasta, Nogueras-Iso e Zarazapa-Soria (2010) apresentam uma nova noção de ontologia na Linguística, as ontologias terminológicas. Os autores ressaltam que esse tipo de ontologia é muito utilizado em bibliotecas, arquivos, museus ou qualquer outro tipo de registro de informação, servindo para facilitar a busca de recursos armazenados, como na classificação e recuperação de informações. Nessa abordagem, a ontologia terminológica pode ser compreendida como um modelo de domínio ou aplicação, que contém a terminologia necessária de uma área de conhecimento, sendo útil para lidar com questões relativas à ambiguidade entre termos. Pensando nisso, uma ontologia terminológica pode fornecer uma estrutura de inter-relações e conteúdo semântico necessário para definir melhor os termos da área.

Um ponto de atenção nesse tipo de arquitetura condiz com o problema de heterogeneidade na estrutura da ontologia. Quando nos deparamos com uma estrutura com diferentes visões e conteúdos, modelar uma ontologia requer cuidados tanto conceituais como linguísticos. No nosso caso, por exemplo, se considerarmos o sistema ERP, que apresenta a estrutura de diversos departamentos de uma empresa, como as áreas de Recursos Humanos, Finanças, Logística, Compra e Vendas, Marketing, dentre outros, integrar todas as áreas e sua terminologia é um desafio, especialmente por cada setor ser organizado de uma forma, e os termos apresentarem variações conceituais e lexicais. A princípio, a terminologia deveria ser padrão para que fosse possível utilizá-la

em todos os sistemas terminológicos, mas, como sabemos, isso não é algo tão simples. Dessa forma, Lacasta, Nogueras-Iso e Zarazapa-Soria (2010) indicam que seria necessário harmonizar as ontologias para um formato comum de representação, permitindo que toda a estrutura, conceitos, termos e suas relações, fossem compartilhadas. Essa problemática apresentada pelos autores é crucial para este estudo, visto que nosso objetivo é trabalhar na harmonização de termos e conceitos, considerando os aspectos semântico-terminológicos. Essa lacuna apontada por Lacasta, Nogueras-Iso e Zarazapa-Soria (2010) evidencia a necessidade de se trabalhar com as questões atinentes à harmonização tanto no contexto linguístico/terminológico como no computacional.

Similar ao formato da ontologia, os grafos de conhecimento (*Knowledge Graphs* - KGs) são uma forma de representar o conhecimento por meio de nós e arestas. Esse recurso é utilizado tanto na área da Matemática quanto na das Ciências Computacionais, mas aparecem timidamente em trabalhos na área da Linguística. No campo da Matemática, a primeira aplicação ligada à teoria dos Grafos remete ao século XVIII, quando o matemático Leonahrd Euler utilizou tal princípio para calcular a distância que o imperador da Prússia deveria andar pela cidade de Königsberg, assegurando que o soberano passasse apenas uma vez por cada uma das sete pontes da cidade (cf. Barrasa *et al.*, 2022; Kejriwal *et al.*, 2021). Pela perspectiva de Euler (1735?), cada parte da cidade poderia ser considerada como um nó e cada ponte representaria as relações entre as áreas do local.

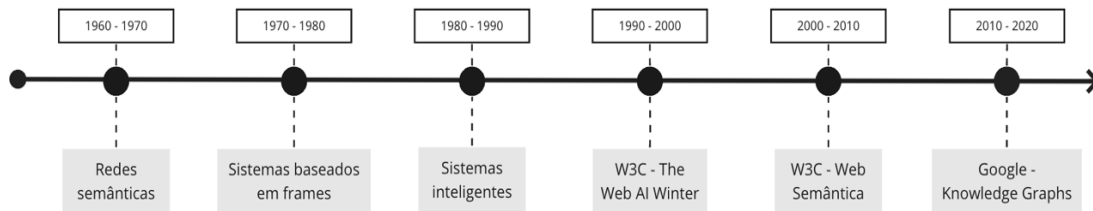
Em seu formato original, os grafos de conhecimento originaram-se das redes semânticas e, atualmente, estão ganhando mais espaço nas empresas de tecnologia e informação no que tange à representação do conhecimento (Gomez-Perez; Pan; Vetere; Wu, 2017; Kejriwal, 2021). Embora essa estrutura já tenha sido utilizada para solucionar problemas matemáticos no século XVIII e seja derivada das redes semânticas, foi apenas com a publicação do artigo “*Introducing the Knowledge Graph: things, not strings*”<sup>38</sup> (Singhal, 2012), pela Google, em 2012, que essas redes foram realmente reconhecidas na área computacional. Na área de Inteligência Artificial, os KGs apareciam em sistemas inteligentes (baseados em regras). Com o anúncio da Google sobre os *Knowledge Graphs*, voltou-se a utilizar tal noção nos trabalhos de

---

<sup>38</sup> SINGHAL, Amit. Introducing the Knowledge Graph: Things, Not Strings. *In.*: Blog Oficial da Google. Califórnia, 16 maio 2012. Disponível em: <https://blog.google/products/search/introducing-knowledge-graph-things-not/>

PLN. Martin, Szekely e Allemang (2021) traçam um histórico do uso de grafos, desde as redes semânticas até os KGs, conforme Figura 30.

Figura 30 - Linha do tempo da representação do conhecimento



Fonte: Adaptado de Martin, Szekely e Allemang (2021, p. 5).

Na Figura 30, podemos observar que a representação do conhecimento parte das redes semânticas, definindo muitas características formais para a organização e estruturação do conhecimento. Posteriormente, entre os anos 70 e 80, busca-se registrar mais detalhadamente o conhecimento, adicionando atributos e explorando diferentes formas de organizá-lo, como, por exemplo, os *frames* de Minsky. Mais adiante, a representação do conhecimento é utilizada para estruturar os sistemas inteligentes, trazendo novos desafios na forma de arquitetar o conhecimento. Com o advento da Internet e com a revolução dos dados, eles deixam de ser armazenados em centros específicos e passam a ser descentralizados com acesso à Web (Martin; Szekely; Allemang, 2021). Assim, o Consórcio W3C busca desenvolver diretrizes para representar o conhecimento na Internet, a fim de que a representação tenha um padrão. Nos anos 2000, com a Web Semântica em pleno desenvolvimento, elabora-se uma linguagem para representar o conhecimento que seja legível por todos e, finalmente, entre os anos de 2010 e 2020, a Google lança os *Knowledge Graphs* como uma forma mais arrojada para tratar da representação do conhecimento.

A robustez dos *Knowledge Graphs* pode ser observada, por exemplo, no âmbito empresarial, visto que eles são ideais para realizar buscas na Internet e servem, também, como base para muitos outros sistemas e aplicações, como, por exemplo, o gerenciamento de informações de empresas. Para Gomez-Perez, Pan, Vetere e Wu (2017), a constante busca pelo uso de KG se dá pela interligação dos dados não apenas na Web, mas também em sistemas complexos de aplicativos específicos de combinação de dados corporativos. Para trabalhos que envolvem o desenvolvimento de estruturas representacionais de conhecimento de produtos em empresas, os grafos de conhecimento – *Knowledge Graphs* (KGs) – são uma boa opção para se trabalhar com

operações de pesquisa e relacionamentos e com dados mal formalizados, mas altamente interconectados.

Como tratamos das ontologias e considerando a relevância de se usar *Knowledge Graphs* no contexto empresarial, principalmente se levarmos em conta a estrutura do ERP, é importante discutir a relação entre as duas formas de representação do conhecimento. De acordo com Steth, Gyrard e Padhee (2019), o KG é uma forma de representar o conhecimento e pode ser referido como redes semânticas, grafos conceptuais e ontologias. Ambas as estruturas, ontologias e KGs, apresentam o conhecimento de forma muito semelhante. Segundo Kejriwal e Knoblock (2021), a estrutura de um *Knowledge Graph* é uma ontologia, por isso há essa aproximação entre os conceitos.

Pan e Wu (2017) concebem a entidade como a unidade básica de um *Knowledge Graph*, a qual pode conter diversos atributos e são conectadas por relações. Tais relações podem ser também usadas como pontes entre dois grafos separados, criando *interlinks* entre eles. Na visão de Liu e Han (2018), o KG pode ser visto como um conjunto de dados que organiza o conhecimento humano de forma estruturada. Nesse sentido, o conhecimento é representado com entidades concretas e conceitos abstratos, os quais apresentam multirrelações entre eles, definidos em alguns dicionários legíveis por máquina, como as ontologias.

No que tange à linguagem de um *Knowledge Graph*, é importante destacar que o esquema de linguagem de entidades e relações se relaciona à ontologia. Nesse sentido, a ontologia permite descrever e representar toda a informação presente em um KG (Pan; Wu, 2017). À vista disso, a ontologia apresenta dois padrões de representação: RDF e OWL. O padrão RDF (*Resource Description Framework*) é recomendado pela W3C e consiste em um esquema de linguagem simples, permitindo que se declarem classes/propriedades utilizando a classe como nível de linguagem pré-definida. Ademais, RDFs podem especificar algumas dependências entre classes e propriedades. Assim, o padrão RDF é composto por um conjunto de triplas, ou seja, nó – arco – nó interligados. Embora tal esquema de linguagem seja moderno para a representação de *Knowledge Graphs*, podemos identificar algumas limitações nessa linguagem, como, por exemplo, não suportam negações, nem o alinhamento no nível da instância e não fornecem a definição do conceito (Kejriwal; Knoblock, 2021; Pan; Wu, 2017).

Tendo em vista as limitações do padrão RDF, o padrão OWL surge para suprir tais restrições, gerando um esquema de linguagem mais compreensível. A linguagem

OWL (*Web Ontology Language*), que também é recomendada pela W3C, tem como base uma descrição lógica e segue uma semântica mais formal, incluindo em seu esquema as linguagens RDF, RDFs e SPARQL. De acordo com Pan e Wu (2017), o esquema OWL pode ser utilizado para checar inconsistências lógicas e semânticas nos *Knowledge Graphs*, além de suportar consultas de perguntas e respostas (Q&A).

O grande diferencial dos *Knowledge Graphs* para as outras arquiteturas reside na possibilidade de conectar dados interligados e de diferentes bases. De acordo com Martin, Szekely e Allemang (2021), atualmente, no campo de gestão de negócios, utilizar KGs é uma solução para o *Data Fabric*, isto é, uma arquitetura de dados da empresa distribuídos e dinâmicos, possibilitando que diferentes áreas do negócio possam gerar seus dados próprios e torná-los reutilizáveis para outros setores da instituição. Ao unir *Knowledge Graphs* e *Data Fabric*, promovemos a aproximação entre gerenciamento de dados e de conhecimento de forma mais eficiente.

Atualmente, os *Knowledge Graphs* atraem áreas como *Representation Learning* – subárea do *Deep Learning* – e Processamento de Linguagem Natural, especialmente por conseguirem capturar o contexto, prezar pela qualidade e pela validade da estrutura do conhecimento e permitir a extração de conhecimento de domínio específico. Além disso, esse tipo de arquitetura permite novas elaborações de aplicações para pesquisa, recomendação, personalização, propaganda tanto no âmbito da Internet como no de negócios corporativos (Steth; Gyrard; Padhee, 2019).

Sendo nosso objetivo, em nível prático, representar a estrutura do ERP – sistema de planejamento de recursos empresariais, que auxilia na administração de toda empresa –, optamos por representar o conhecimento do sistema utilizando *Knowledge Graphs*, já que essa arquitetura permite combinar dados de empresas e conhecimento, bem como pela forma como ela consegue capturar o contexto do conhecimento. Outro ponto importante pelo qual optamos pelos KGs diz respeito à interrelação entre dois ou mais *Knowledge Graphs* e o *Data Fabric*, uma vez que, em empresas, geralmente, o *software* ERP abrange diversos departamentos, os quais compartilham uma gama de dados.

Pensando nesse contexto empresarial e na estrutura do *Knowledge Graph* para a representação do conhecimento, Martin, Szekely e Allemang (2021) ressaltam um aspecto crucial na construção de KGs em empresas: a harmonização de dados. Grandes empresas têm muitos conjuntos de dados, e isso pode gerar problemas, sobretudo com dados semelhantes, mas de diferentes setores do negócio. Uma forma de solucionar esse impasse é buscar harmonizar os dados institucionais, visto que a representação do

conhecimento da empresa pode contribuir para lidar com a desambiguação de termos e conceitos e evitar inconsistências. Para isso, os autores apresentam dois pontos: o primeiro se refere às relações entre entidades, que precisam ser consistentes e precisas, sempre considerando as informações e a estrutura do domínio a ser representado, e o segundo ponto se relaciona à terminologia. A questão principal apresentada por Martin, Szekely e Allemang (2021) está ligada a como lidar com um termo e seu significado em uma base de dados específica. Além desse questionamento, aqui nos perguntamos: como lidar com diferentes termos, de diferentes áreas, que representam um mesmo conceito em uma base de dados específica?

Para Martin, Szekely e Allemand (2021), a harmonização deve ocorrer sem mudar as fontes de dados referentes aos termos e às relações, mas se deve buscar combiná-los com a representação do conhecimento. Do nosso ponto de vista, além da perspectiva e metodologia de Martin, Szekely e Allemang (2021), adicionamos uma etapa prévia: deve-se explorar e examinar os fenômenos linguísticos de um *corpus* específico do domínio a ser trabalhado. No nosso caso, o estudo semântico e terminológico tanto para a elaboração da proposta de harmonização como para a criação do KG está fundamentado na análise dos termos em um *corpus* composto por textos das diferentes áreas abrangidas pelo ERP, a fim de extrair e analisar a terminologia dos departamentos, conectando termos e variantes, para, então, trabalhar na metodologia dos autores, combinando os *Knowledge Graphs*, *Frames* e os termos.

Nosso trabalho tem o objetivo de desenvolver um *Knowledge Graph* com *frames* semânticos, que represente a estrutura do *software* ERP, que possa ser utilizado no *prompt* de um LLM e tratar da harmonização de termos e conceitos nessa arquitetura. Após a revisão da literatura acerca da Inteligência Artificial, do Processamento de Linguagem Natural, dos *Large Language Models* e da representação do conhecimento – ontologias, *Knowledge Graphs* e *Frames* – buscamos artigos, dissertações, teses e experimentos que aproximassem as abordagens atinentes aos pontos principais desta tese. Para isso, em um primeiro momento, abordaremos os trabalhos relacionados aos *frames* semânticos e, depois, discutiremos os trabalhos atinentes à harmonização de termos e conceitos.

No que tange aos estudos relacionados à elaboração e ao uso de representações do conhecimento, como ontologia, destacamos alguns realizados na interface linguístico-computacional. Vieira *et al.* (2005) elaboram um tutorial para discorrer sobre a Web Semântica, abordando o conceito de ontologia, lógica descritiva e mecanismos

de raciocínio. Outro trabalho nesse âmbito foi desenvolvido por Trojahn, Quaresma e Vieira (2008), que propõem um processo para o mapeamento de ontologias multilíngues de descrição lógica. Para isso, a ontologia na língua-fonte é traduzida para a ontologia na língua-alvo a partir do uso de base de dados lexicais ou dicionários. Em seguida, a ontologia traduzida é utilizada como *input* para o processo de mapeamento. No próximo passo, o mapeamento é computadorizado por agentes especializados diferentes abordagens. Depois disso, os agentes usam argumentação para mudar os resultados locais, visando à concordância acerca do mapeamento. Além dessa pesquisa, Lopes *et al.* (2009) evidenciam o uso de uma ferramenta para a construção de ontologia a partir de um experimento na área da saúde. Para essa finalidade, inicialmente, extraíram-se termos específicos do domínio por meio de um *corpus* e se compararam os resultados com os gerados pela ferramenta ontológica. Por fim, concluiu-se que o trabalho na interface linguístico-computacional apresentou bons resultados, principalmente pela inserção de informações sintáticas e semânticas com a ontologia.

Vieira e Lopes (2010) apresentam os benefícios de se utilizarem as técnicas e as abordagens de Processamento de Linguagem Natural para o tratamento de linguagens específicas. Destacam-se as discussões acerca de procedimentos utilizados para suprir os desafios da Web Semântica, como a necessidade de representar formalmente um determinado domínio por meio de ontologias. Mais tarde, Wendt e Vieira (2011) tratam da extração de contexto definatório em textos em língua portuguesa. Os autores partem do princípio de que algumas tecnologias desenvolvidas se fundamentam na Linguística para dar suporte a outras tarefas, como a busca de informações, geração de sumários e elaboração de dicionários e glossários. Para a compilação de dicionários, uma das etapas principais é a identificação dos conceitos de um domínio e de sua descrição. Tal descrição, que é extraída de textos, é denominada de contexto definatório. Essa etapa, geralmente, é feita por especialistas do domínio e, para tornar essa extração mais eficiente e menos custosa, busca-se automatizar tal procedimento a partir da construção de ontologias.

Outro estudo sobre representação do conhecimento foi desenvolvido por Granada *et al.* (2012), que propõem uma metodologia para a elaboração de um *corpus* comparável multilíngue, utilizando ontologias multilíngues de domínios específicos. Para isso, utilizaram-se ontologias de dependência de múltiplos domínios, proporcionando meios de elaborar um *corpus* de domínio específico da *web*, haja vista que as relações e os conceitos provenientes das ontologias permitem que se extraiam os



aspectos semânticos e contextuais de um dado domínio. Por fim, Freitas, Bordini e Vieira (2017) apresentam uma proposta para o desenvolvimento de sistemas multiagentes como modelos ontológicos que auxiliam na geração de código. Para tanto, propõe-se o desenvolvimento de um sistema baseado em ontologia, utilizando uma linguagem geral para apresentar e especificar um sistema multiagente, resultando, assim, na integração de seis diferentes aspectos e servindo como base para a geração de código em JaCaMo, abordagem orientada para programação de multi-agente.

Do ponto de vista computacional, os *frames* semânticos apresentam diferentes formas de serem utilizados e trabalhados em recursos de PLN. No que concerne ao uso de *frames* e ontologias, Bārzdiņš *et al.* (2008) se propõem a criar uma representação da FrameNet em formato de uma ontologia multidimensional 4D. Embora a FrameNet seja um recurso lexicográfico que utiliza *frames* semânticos para a sua implementação, apresentando uma visão distinta da semântica lexical e a sua interação com a estrutura sintática das sentenças e a conexão entre os *frames* e o sentido das palavras, a elaboração e a implementação dos *frames* apresenta uma limitação quanto a sua formalização, sobretudo no nível da descrição. Justamente por isso é difícil realizar uma análise semântica sem a supervisão humana. A partir da criação de *frames* como ontologias multidimensionais, seria possível capturar o conhecimento codificado nos textos anotados manualmente e, posteriormente, essa metodologia serviria para converter textos anotados da FrameNet para representação ontológica multidimensional. Em um outro estudo que aproxima os *frames* das ontologias, Tan, Kaliyaperumal e Benis (2011) buscam construir um léxico baseado em *frames* para o domínio biomédico por meio de ontologias. Eles utilizam a FrameNet para o processo de anotação de papéis semânticos, uma vez que construir um *corpus* e anotá-lo requer muito tempo. Considerando que a FrameNet não apresenta muitos *frames* relacionados à área biomédica, os autores propõem construir ontologias do domínio para suprir tal lacuna e implementá-las junto com os *frames*. Seguindo a estrutura da FrameNet, com unidades lexicais e elementos de *frames*, os autores destacam que as ontologias podem ser usadas como descrições de *frames* semânticos em um léxico. Para isso, eles, inicialmente, definem quais pontos serão abarcados pelos *frames* e suas relações; em seguida, descrevem os cenários e as relações por meio dos termos e, finalmente, utilizam os termos ontológicos coletados de um *corpus*, que fornecem uma lista de predicados específicos.

Moreira e Salomão (2012) também tratam da questão da ontologia com *frames*. Em seu estudo, as autoras apresentam um processo de desenvolvimento de uma FrameNet que inclui a análise ontológica, permitindo que melhor se estabeleçam as relações entre *frames*. A priori, os *frames* são incorporados à base, obedecendo a critérios semiformais, bem como a alguns critérios ontológicos. A posteriori, com a análise de sentenças do *Corpus* do Português, alinham-se os elementos de *frames* com os conceitos pertencentes às ontologias de topo. Após essa etapa, detectam-se as sentenças que apresentam o sentido prototípico da FrameNet. Com esse trabalho, as autoras apontam que tal metodologia pode facilitar a formalização das relações entre *frames* distintos e polissêmicos.

Com relação ao uso de *frames* em outras aplicações de PLN, Agarwal, Bauer e Rambow (2014) utilizam os *frames* para tratar da extração de relações sociais de textos narrativos, assim como a gerar figuras tridimensionais de descrição textual. Para a primeira tarefa, os autores observaram que os *frames* não foram suficientes, visto que eles não têm uma cobertura lexical completa e as sentenças são coletadas de um *corpus* pequeno. Devido a isso, foram utilizados *parser* e árvores semânticas para treinar o modelo. Para a tarefa de gerar imagens, os autores destacaram que os *frames* são um ótimo ponto de partida no que se refere à representação do conhecimento para uma tarefa específica. Por se tratar de um grande repositório de *frames*, incluindo anotações e relações, é possível gerar imagens a partir dos dados da FrameNet. Por fim, os autores salientam que é preciso se aprofundar em algumas limitações, como a falta de cobertura completa dos itens lexicais e os dados anotados serem de um *corpus* pequeno, pontos que dificultam a implementação com técnicas de *Machine Learning*, para que os *frames* da FrameNet possam ser mais bem utilizados.

Outra possibilidade é apresentada por Mousselly-Sergieh e Gurevych (2016), que tratam do enriquecimento da Wikidata a partir do uso de *frames* semânticos. Nesse sentido, eles alinham as propriedades da Wikidata com os *frames* e recursos da FrameNet por meio de *embeddings* para que as palavras semelhantes estejam próximas, uma vez que elas evocam os mesmos *frames*. Uma perspectiva diferente para se trabalhar com *frames* no campo computacional foi desenvolvido por Corcoglioniti, Rospeoche e Aprosio (2016), que propuseram uma abordagem para atualizar as informações de uma ontologia da FrameBase, base derivada da FrameNet, a partir da extração de triplas de RDF de textos em inglês. Com o uso do sistema PIKES, que extrai o conhecimento de textos de linguagem natural baseado em *frames*, é possível

coletar instâncias e relações complexas entre elas e identificar os *frames* nos textos em inglês. Desse modo, observou-se que o PIKES consegue lidar com a atualização de uma ontologia da FrameBase, detectar *frames* de acordo com a FrameNet e extrair *Knowledge Graphs* de grandes corpora.

Recupero, Alam, Gangemi e Presutti (2017) buscaram desenvolver um novo algoritmo, o *TakeFive*, que realiza a anotação de *frames* semânticos, utilizando *frames* orientados por *Knowledge Graphs* gerados pelo Framester. Os autores salientam que os *Knowledge Graphs* são limitados quanto à apresentação de informações situacionais e contextuais e, para suprimir essa lacuna, os *frames* contribuem com informações extras.

Botschen (2019) propõe-se a examinar o significado da linguagem capturado pelo conhecimento de mundo, considerando (i) o conhecimento sobre situações e ações expressas em textos e (ii) o conhecimento relacional estruturado armazenado em bases de conhecimento. Os dois ramos podem ser estudados por meio da representação de *embeddings*. Tendo em vista que há diferentes formas de representar os *embeddings*, o desafio principal é determinar qual tipo de *embedding* é melhor para compreender a linguagem automaticamente considerando os dois ramos. Para isso, a autora utiliza a FrameNet, já que o recurso fornece estruturas acerca do conhecimento e seus significados, bem como as relações entre *frames* e predições. Bostchen (2019) ressalta que o uso de *frames* nesse tipo de estudo é muito benéfico na medida em que auxilia na compreensão de frases curtas e ambíguas.

No que se refere ao uso de *frames* para a análise discursiva computacional, Minnema *et al.* (2021) examinaram o discurso midiático italiano acerca do feminicídio, utilizando *frames* da FrameNet. Os autores indicam que muitos estudos propostos pela comunidade linguística/computacional italiana focam no uso de FrameNets multilíngues, mas poucos estudos exploram a usabilidade dos *frames* na prática. Para isso, utilizou-se um *dataset* de notícias sobre feminicídio, em seguida, o LOME *parser* para a anotação dos *frames* e, para a fase de testes, usou-se o *dataset* EVALITA 2011. Após os experimentos, os autores observaram que os *frames* contribuem na identificação de construções linguísticas que evidenciam os agressores, além de demonstrarem que os *frames* são importantes para analisar outras questões sociais.

Outra perspectiva para análise linguística e computacional com *frames* foi proposta por Dey e Girju (2022). As autoras apresentam uma abordagem computacional e análise linguística relativas à linguagem empática. Para esse estudo, utilizou-se um *corpus* composto por 440 textos de narrativas simulando uma interação médico-paciente

de alunos de Medicina. Com isso, realizou-se uma análise das palavras relacionadas à empatia a partir das perspectivas cognitiva, afetiva e social. Para a análise, recorreu-se à FrameNet para identificar os *frames* em que as palavras relacionadas à empatia se encaixavam e utilizaram-se redes neurais recorrentes e modelos *Transformers* para classificar as diferentes formas de empatia. Tal estudo evidenciou que as informações semânticas contidas no nível da sentença podem ser muito úteis para identificar os termos relativos à empatia e fornecer detalhes dos diferentes tipos de empatia.

Acerca do desenvolvimento de recursos para a criação de *frames*, Koeva e Doychev (2022) apresentam a ferramenta BullFrame, elaborada para criar, editar, validar e ver *frames* conceituais. Nesse estudo, os autores indicam que a grande diferença entre os frames provenientes da FrameNet e da BullFrame está na forma de representar e definir os dados. Enquanto na FrameNet os *frames* são definidos por um conjunto de substantivos que podem combinar com um determinado verbo, na BullFrame isso ocorre por meio da representação ontológica das classes semânticas de um substantivo, conectando FrameNets e WordNets.

Sasano (2022) se propõe a elaborar um recurso que gere automaticamente *frames* conectados por diferentes línguas de FrameNets. Para isso, ele utiliza os *frames* da FrameNet em japonês e, por meio de *embeddings* de palavras entre línguas, associa com os *frames* da FrameNet em inglês. Essa abordagem ocorre em duas etapas. A primeira consiste em extrair os *frames* candidatos da FrameNet em japonês, considerando as palavras evocadoras do *frame* para a FrameNet em inglês. O segundo passo é alinhar os *frames* considerando os elementos de *frame*. Esse estudo aponta para a possibilidade de se gerarem *frames* de forma automática, sobretudo no contexto translacional. Contudo, sabe-se que, para gerar *frames* em diferentes línguas, deve-se considerar o contexto social e cultural do par linguístico, e isso se refletirá na organização da estrutura.

Ainda sobre o desenvolvimento de ferramentas para elaborar *frames*, destacamos o Lutma (Torrent *et al.*, 2022), recurso que permite criar *frames*, contribuindo para a sua consistência e promovendo a integração de *frames* desenvolvidos pela comunidade. A ferramenta auxilia no desenvolvimento de *frames* com o objetivo de enriquecer a FrameNet com diferentes *frames*, perspectivas e línguas. Tendo em vista que, atualmente, em PLN, tem-se buscado padronizar as tarefas e evitar a ambiguidade em diferentes tipos de anotação, a construção do Lutma serve para que haja uma consistência no formato de elaboração da estrutura semântica. No entanto, é importante

destacar que os *frames* produzidos no Lutma ficam disponíveis para todos os usuários da ferramenta, visto que o objetivo do recurso é compartilhar com toda a comunidade diferentes estruturas. Para os casos em que a estrutura a ser elaborada está ligada a dados sigilosos, como ocorre, por exemplo, em muitas pesquisas na área corporativa, tecnológica e médica, utilizar o Lutma não é uma boa opção, já que não se pode compartilhar os dados de pesquisa com toda a comunidade.

Por fim, no que se refere aos estudos que utilizam *frames* na Computação, Giacomini e Schäfer (2020) desenvolvem e implementam um método para modelagem de variação terminológica baseada em ontologias e *frames* para textos relativos a produtos técnicos. Os autores partem do conceito terminológico wüsteriano para tratar da padronização e, em seguida, baseiam-se nos pressupostos de Faber para tratar da Terminologia baseada em *Frames*. Para a modelagem, utilizaram-se os *frames* da FrameNet para identificar os tópicos a serem trabalhados nos textos específicos do domínio e, para as ontologias, optou-se por ontologias no formato OWL de recursos já existentes, como SUMO e DOLCE. No que tange à anotação, alguns conjuntos de termos extraídos são adicionados aos *frames*, criando um léxico de conjunto de associações. Após a tokenização do léxico, os *tokens* são anotados com etiquetas de *frames* e servem como insumo para o léxico. Por fim, aplicaram-se restrições ontológicas às combinações de elementos de *frames* e restrições sintáticas na extração de termos e variantes. Os autores destacam que esse método computacional é promissor e aplicável a domínios que se assemelham ao trabalho desenvolvido por eles.

Com relação aos estudos e trabalhos voltados para a harmonização, Pardo *et al.* (2011) propõem uma harmonização de modelos de programas por meio de ontologias, uma vez que as empresas utilizam diferentes programas tanto internamente como externamente. A questão principal discutida pelos autores diz respeito à busca de uma harmonização de modelos e técnicas, considerando as diferenças e inconsistências terminológicas que cada modelo apresenta. Por não haver uma terminologia consistente sobre métodos e técnicas para harmonizar, não há um consenso na aplicação e no desenvolvimento dessa tarefa. Uma alternativa para esse problema é utilizar ontologias para a harmonização, pois elas permitem o compartilhamento de conhecimento, por meio de vocabulário, termos, conceitos e relacionamento especializados. Dessa forma, é possível eliminar problemas de harmonização e conflitos terminológicos existentes nos modelos, métodos e técnicas.

Outro estudo acerca da harmonização foi desenvolvido por Bergovec e Runjaić (2012), que analisaram a harmonização a ser aplicada na base de termos Struna, uma base terminológica Croata para a padronização da língua. A ideia principal para criar a base de termos e padronizá-la surgiu, sobretudo, da necessidade de reavaliar a linguagem devido à globalização e à necessidade de se comunicar em inglês. Para a tarefa de harmonização, foi considerada a unificação de conceitos e sistemas de conceitos, definição de termos, redução de homonímia, eliminação de sinonímia, definição de designações, inclusão de abreviaturas e símbolos e criação de novos termos. O estudo seguiu os padrões estabelecidos pela ISO para harmonização, bem como os preceitos de Termmerman (2004) para a análise terminológica. Os autores destacam que a harmonização da base de termos foi demorada devido à necessidade de se aplicarem diferentes metodologias para termos homônimos e para termos sinônimos.

Ainda que o escopo do trabalho desenvolvido por Cram e Daille (2016) não seja a harmonização de termos e conceitos *per se*, os autores apresentam a ferramenta *TermSuite*, utilizada para a extração terminológica a partir dos preceitos da *Term Extraction Tools*. Um dos pontos de destaque da ferramenta é a possibilidade de detectar variantes. Para isso, eles incorporam um componente chamado de reconhecimento de termo variante, capaz de melhorar a extração de termos, tornando os candidatos a termos ranqueados mais precisos e, conseqüentemente, melhorando a detecção terminológica.

Atentando para as constantes mudanças terminológicas na área da Biologia, Grenié *et al.* (2021) tratam da harmonização de dados relativos à biodiversidade. O foco principal do trabalho é tornar os dados disponíveis considerando a qualidade, a interoperabilidade e a redundância. Os autores apontam que muitas ferramentas foram criadas para a tarefa de harmonização de dados, mas não há um guia nem consenso sobre como combinar as ferramentas de forma significativa e eficiente. Para harmonizar, eles desenvolveram dois passos obrigatórios e dois passos opcionais. De início, é preciso pré-processar os nomes. Essa etapa é importante, pois é possível observar muitas mudanças de estilo de escrita e muitos problemas sintáticos, como, por exemplo, maiúscula/minúscula, abreviações, anotações, autores e nomes híbridos. Logo, para que haja uma padronização da taxonomia, é necessário pré-processar todos os dados. Nesse estágio, os autores indicam um passo opcional: contrastar as correspondências exatas com diferentes bases de dados específicas da área, a fim de averiguar diferenças entre elas. O segundo passo consiste em checar os dados em bases

de dados específicas para verificar a cobertura dos dados. Por fim, o último passo adicional é contrastar os termos sem correspondentes com as correspondências difusas, procurando encontrar erros de ortografia ou outros problemas que possam interferir na correspondência exata.

O estudo de Adhikari *et al.* (2021) propõe uma harmonização de *datasets* de diferentes estudos. Para isso, eles partem do conceito de *data pooling*, isto é, um processo em que diferentes *datasets* são combinados e, assim, obtêm-se os dados a serem harmonizados. Para os autores, a harmonização de dados se refere aos esforços fornecidos para se encontrar a compatibilidade de *datasets* de fontes heterogêneas e, dessa forma, permitir combinar ou integrar os dados de forma coerente. Por essa perspectiva, a harmonização pode considerar duas abordagens: a prospectiva, referente à harmonização que ocorre no estágio inicial do estudo ou antes da coleta de dados (nessa abordagem, os pesquisadores definem as medidas e as variáveis para a coleta e padronizam a operação); e a retrospectiva, uma abordagem mais flexível, que objetiva sintetizar as informações já coletadas. Nessa abordagem, os pesquisadores definem as variáveis e avaliam a compatibilidade das informações coletadas e o potencial para criar variáveis únicas para a harmonização. Para seu experimento, foram utilizados *datasets* da área da saúde, seguindo uma harmonização pelo viés retrospectivo. Conforme Adhikari *et al.* (2021) observaram, a harmonização de dados e as estratégias utilizadas contribuíram para a criação de um *dataset* comparável e para o *data pooling*.

Seguindo a mesma perspectiva de harmonização de dados, Kumar *et al.* (2021) fizeram uma revisão da literatura relativa à harmonização de dados para *datasets* heterogêneos. Os autores coletaram artigos e publicações que tratavam sobre técnicas, modelos, ferramentas e algoritmos utilizados para a harmonização. No contexto do *Big Data*, a produção de dados é massiva, sobretudo nas áreas da saúde, telecomunicação, finanças, atacado, justiça, marketing, energia, seguros, educação, agricultura e planejamento urbano, e não há uma padronização quanto ao gerenciamento desses dados. À vista disso, trabalhar na harmonização de dados contribui para a descrição da representação uniforme de dados heterogêneos. Kumar *et al.* (2021) salientam que poucos estudos tratam da harmonização de dados com dados textuais, técnicas e métricas de performance. No levantamento bibliográfico dos autores, foram encontrados 25 estudos que abordavam a harmonização para a solução de problemas de heterogeneidade. Dentre as técnicas utilizadas, algumas abordagens utilizaram a Web Semântica e *BD Tools*, indicando que a performance de precisão, *recall* e *f-score* foram

melhores do que as técnicas já existentes. Além disso, outros estudos utilizaram diferentes abordagens, mas nenhum deles propôs a criação de valor e métrica para a harmonização. É interessante destacar que, na área médica, alguns estudos evidenciaram a relevância de se utilizarem técnicas de harmonização com *Machine Learning*, *Deep Learning* e ontologias, fornecendo mais informações para essa tarefa.

Pelo viés da harmonização em dados da saúde, Bergman *et al.* (2022) discutem sobre a necessidade de se padronizarem e harmonizarem nomes de medicamentos. Essa discussão surgiu em 2020, quando a União Europeia debateu acerca da relevância de agências médicas europeias apresentarem e desenvolverem uma harmonização para os nomes dos medicamentos e, posteriormente, disponibilizarem as informações do produto em formato eletrônico. Nesse contexto, os autores realizaram um estudo com o *corpus* EMA PI e com o *Transformer* BERT para buscar as relações de similaridade entre os nomes dos remédios e as informações em sentenças com o objetivo de padronizar os dados. Os experimentos indicaram que a metodologia utilizada foi significativa para a padronização dos dados a partir da identificação da similaridade semântica entre os nomes dos medicamentos.

Outro estudo buscando harmonizar dados na área médica foi desenvolvido por Urbanowicz *et al.* (2022), que propuseram uma abordagem semiautomática de harmonização de termos da área médica a partir de ensaios clínicos de hipertensão arterial pulmonar. O *pipeline* consistia em (i) pré-processar os dados, (ii) encontrar correspondentes exatos, (iii) encontrar correspondentes difusos, (iv) integrar os dados e (v) realizar o controle de qualidade das anotações feitas por especialistas. É interessante destacar que a etapa de avaliação da qualidade da anotação feita por especialistas é o diferencial do trabalho, uma vez que essa etapa contribuiu para resolver problemas de ambiguidade no mapeamento de termos e proporcionou uma maior precisão e acurácia dos dados.

Por fim, apresentamos o trabalho de Yamagata *et al.* (2022), que desenvolveram uma forma de harmonizar as ontologias relativas à COVID-19, visto que, na época da pandemia, muitos dados foram publicados e não havia uma padronização quanto à disseminação de informação. Nesse sentido, os autores buscaram as sete maiores ontologias atinentes à COVID para trabalhar na harmonização. A primeira etapa consistiu em analisar a sobreposição de termos das ontologias, muitas delas provenientes da *The Open Biological and Biomedical Ontologies* (OBO Foundry), grupo dedicado a construir e manter ontologias referentes à área da Saúde. O próximo



passo foi alinhar as ontologias, seguindo o formato de *Basic Formal Ontology* (BFO). Os autores destacam que essa abordagem para a harmonização de ontologias auxilia na identificação de sobreposição de ontologias e, posteriormente, é necessário analisar e comparar as entidades entre as arquiteturas para realmente haver uma harmonização e diminuir a heterogeneidade e as diversas granularidades.

Em vista do que já foi desenvolvido na área computacional relativo à harmonização, à representação do conhecimento e ao uso de LLMs na área terminológica, para este trabalho, buscaremos subsídios nos *Knowledge Graphs*, como afirmamos, juntamente com os *frames* semânticos para construir a arquitetura do ERP. Dado o contexto empresarial em que esta tese se insere, trabalhar com KGs se torna fundamental, sobretudo pela possibilidade de que tal estrutura fornece o acesso a dados de diversos setores interconectados. No caso da abordagem semântico-terminológica de harmonização que estamos propondo e, considerando os dados a serem harmonizados, não trabalharemos com informações de apenas um departamento do ERP, mas com termos de diferentes áreas e de diferentes bases de dados. Nesse sentido, os *Knowledge Graphs* são vantajosos para esse tipo de tarefa.

Outro ponto que nos leva a utilizar os *Knowledge Graphs* neste trabalho diz respeito às estratégias e às técnicas a serem adotadas durante no uso de LLMs. Como vimos, utilizar alguma arquitetura de representação do conhecimento no modelo de linguagem contribui para uma maior precisão dos dados e resultados. Após a elaboração da arquitetura do KG e *frames* do ERP, podemos utilizar essa estrutura para fornecer uma instrução mais completa e detalhada e, conseqüentemente, obter uma resposta mais precisa e próxima do que se espera.

Ainda sobre o uso de LLMs, é interessante destacar que grande parte das pesquisas que investigam o uso e a performance dos diferentes grandes modelos de linguagem se voltam para a área da tradução. Muitas pesquisas focam apenas em medir o desempenho do LLM do ponto de vista computacional, alguns trabalhos são desenvolvidos na interface linguístico-computacional. Contudo, poucos estudos são elaborados apenas por linguistas, principalmente em áreas que não são tão ligadas à Tradução.

Durante os estudos que fizemos sobre LLMs, observamos que os aspectos terminológicos são tratados de forma tímida dentro do contexto tradutório, sendo discutidos apenas na análise da tradução de um termo específico de área, a fim de avaliar se o LLM fez uma boa transladação ou não. Ademais, na área da tradução, muito

se discute sobre o uso de *chatbot*, como, por exemplo, o ChatGPT, atrelado à *CAT tool* para traduzir e, depois, o especialista, geralmente o tradutor (e aqui, podemos inserir também linguistas e terminólogos), realiza a pós-edição, conferindo se o texto está coerente e coeso e se as traduções de termos técnicos estão de acordo com o que o cliente espera. Além disso, alguns artigos discutem acerca do uso dos grandes modelos de linguagem na Lexicografia (cf. de Schryver, 2023), abordando o futuro da área na Era dos LLMs e como utilizar os sistemas inteligentes para auxiliar nos trabalhos lexicográficos. Contudo, não encontramos artigos que analisam especificamente termos ou se propõem a trabalhar em tarefas terminológicas, em especial trabalhos desenvolvidos na interface entre Linguística e Computação.

Com relação à harmonização de termos e conceitos, percebemos que muitos trabalhos se aprofundam em harmonizar *datasets*, mas não consideram o ponto de vista linguístico para a elaboração de uma abordagem ou na análise de dados. Isso indica que, para compreender os termos e conceitos, é preciso, também, atentar para os aspectos linguísticos e não somente para a questão computacional. Alguns artigos tentaram buscar soluções para lidar com a ambiguidade ou a similaridade entre termos e textos, mas não se aprofundaram no viés da Linguística para auxiliar nas tarefas. É justamente pensando nessas lacunas e na falta de pesquisas que investiguem o uso de LLMs em tarefas terminológicas, como a harmonização, que nos propomos a elaborar uma abordagem semântico-terminológica para harmonizar termos e conceitos. Em nossa proposição, busca-se construir uma arquitetura que represente o ERP e usá-la no LLM, a fim de automatizar a tarefa de harmonização, bem como mostrar novas possibilidades de se utilizar os grandes modelos de linguagem em outras atividades linguísticas para além da tradução.

Do ponto de vista computacional e considerando o que já foi desenvolvido até o momento, o trabalho de Giacomini e Schäfer (2020), que usam *frames* semânticos para desenvolver e implementar um método para modelagem de variação terminológica baseada em ontologias, aproxima-se mais do que buscamos nesta tese. Em seu estudo, os autores exploram a Terminologia baseada em *Frames*, de Faber (2012), para justificar o uso de *frames* da FrameNet e modelar a terminologia. No entanto, os autores partem de ontologias prontas e trabalharam considerando apenas a parte computacional da tarefa. Para o nosso trabalho, não encontramos uma ontologia bem estruturada do sistema ERP e nem *frames* da FrameNet que tratem dos processos que envolvem os diferentes componentes do *software*. Logo, precisaremos construir o *Knowledge Graph*

e os *frames* do ERP desde o início, considerando os pressupostos linguísticos. Pensando na necessidade de alinhar *frames* e KG, adotaremos a metodologia de Müller (2015) para representar a arquitetura do ERP.

No que se refere à harmonização de termos, o estudo desenvolvido por Bergovec e Runjaić (2012) ainda que se relacione com o que buscamos neste trabalho, não considera algumas questões linguísticas importantes para harmonizar termos e conceitos. Enquanto os autores utilizam a metodologia da ISO para harmonizar os dados e os pressupostos de Termmerman para a análise terminológica, buscamos um direcionamento que incorpora a Semântica de *Frames*. Como já discutimos, seguir apenas a metodologia proposta pela ISO para a harmonização de termos e conceitos não é o suficiente, principalmente por ela não considerar os aspectos social, cognitivo e comunicativo da Terminologia, o contexto e a visão enciclopédica, fatores que permitem compreender melhor as relações entre conceitos e termos e estruturar uma base terminológica.

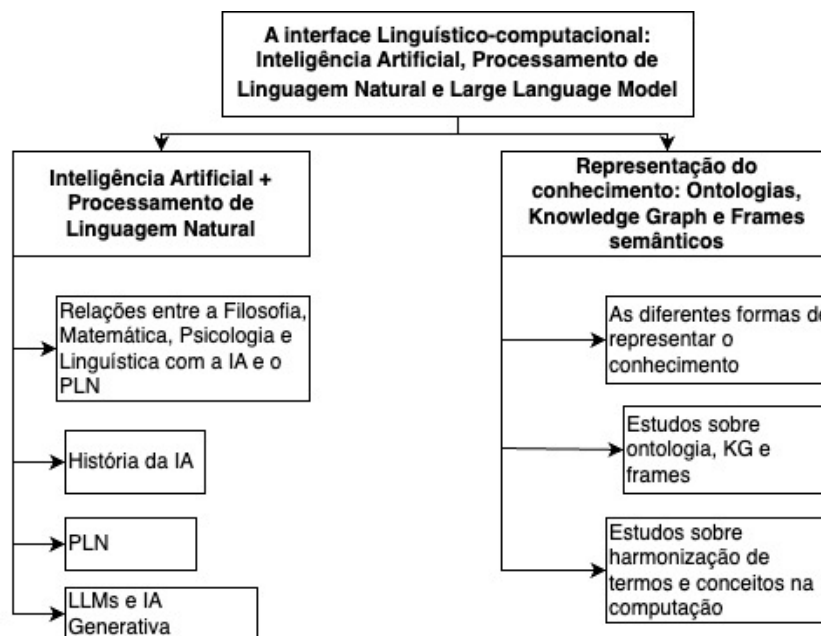
Conforme abordamos previamente, as discussões de Martin, Szekely e Allemang (2021) acerca da necessidade de se harmonizarem dados para a construção de *Knowledge Graphs* se conecta com o que estamos propondo nesta tese. Os autores evidenciam que a harmonização de conjuntos de dados de uma empresa é fundamental para lidar com a desambiguação de termos e conceitos e evitar inconsistências. Ainda que a metodologia apresentada por eles para lidar com um termo e seu significado em uma base de dados não trate de diferentes termos que designam um mesmo conceito, podemos levar em consideração as etapas que os autores propõem para a nossa abordagem.

Martin, Szekely e Allemang (2021) indicam que a harmonização de dados deve acontecer levando em conta as fontes de dados referentes aos termos e que as relações não sejam alteradas. Nesse sentido, deve-se buscar combinar a representação do conhecimento com os dados. Para o nosso trabalho, entendemos que há um passo anterior à metodologia de Martin, Szekely e Allemang (2021). Assim, devemos, primeiramente, explorar os fenômenos linguísticos de um *corpus* especializado do domínio a ser representado. As informações linguísticas coletadas do *corpus* auxiliarão na compreensão e na estruturação do domínio, na identificação de *frames*, termos e variantes. Somente após a exploração do *corpus*, da extração da lista de termos e da criação de *frames* é que podemos combinar a representação do conhecimento com os dados da base terminológica.

Tendo em vista que compilaremos um *corpus* e examinaremos os fenômenos linguísticos tanto para tratar da harmonização como para criar um *Knowledge Graph*, adotaremos a metodologia criada por Müller (2015) para elaborar ontologias com *frames* semânticos, visto que essa abordagem contribuirá para elaborarmos a nossa arquitetura tanto do ponto de vista linguístico como computacional. Como vimos no trabalho de Müller (2015), adicionar as informações dos *frames* em ontologias contribui para uma descrição mais detalhada dos conceitos do domínio. Além disso, utilizar as três abordagens – *Top-down*, *Middle-out* e *Bottom-up* – na etapa de conceitualização da ontologia faz com que a sua estrutura seja mais robusta e as informações provenientes do *corpus* sejam inseridas. Para o nosso *Knowledge Graph*, essa metodologia nos auxiliará na construção do grafo do ERP, considerando a Semântica de *Frames*, assim como na organização e harmonização dos termos e conceitos e, posteriormente, na criação das fichas terminológicas.

Considerando os pontos discutidos neste capítulo, a Figura 31 sinaliza os enfoques que merecem destaque.

Figura 31 - Síntese do capítulo “A interface Linguístico-computacional: Inteligência Artificial, Processamento de Linguagem Natural e *Large Language Model*”



Fonte: Elaborada pela autora.

Como mostra a Figura 31, em um primeiro momento, discutimos sobre a Inteligência Artificial (IA) e o Processamento de Linguagem Natural, apresentando as relações entre a Filosofia, a Matemática, a Psicologia e Linguística com as áreas. Em

seguida, voltamo-nos à história da IA para, então, discorrermos sobre o Processamento de Linguagem natural (PLN). Por fim, nesta seção, abordamos os *Large Language Models* (LLM) e a Inteligência Artificial Generativa. Em um segundo momento, dissertamos sobre a representação do conhecimento, elucidando as diferentes formas de tal representação. Mais adiante, apresentamos estudos sobre ontologia, *Knowledge Graph* e *frames* e, por fim, ocupamo-nos de estudos sobre harmonização de termos e conceitos pelo viés da computação.

Tomando como referência as discussões e abordagens explicitadas até aqui, no próximo capítulo, apresentaremos os procedimentos metodológicos utilizados nesta pesquisa.

## 5 METODOLOGIA

Como nosso objetivo é investigar o papel da Semântica de *Frames* como princípio organizador de uma base de dados ERP, contribuindo para a harmonização terminológica, assim como desenvolver um *Knowledge Graph* com *frames* semânticos do *software* e explorar estratégias e técnicas de *prompt* para trabalhar a harmonização de termos e conceitos utilizando o ChatGPT, neste capítulo, apresentaremos os procedimentos metodológicos adotados para este trabalho. Para isso, na seção 5.1, trataremos da Linguística de *Corpus*, apresentando um panorama dessa abordagem metodológica e discorrendo sobre o planejamento e a compilação dos *corpora* e das ferramentas computacionais utilizadas para a análise. Posteriormente, na seção 5.2, discutiremos acerca dos *Knowledge Graphs* e dos *frames* semânticos, abordando a criação dessa arquitetura e os recursos selecionados para a sua construção. Por fim, a seção 5.3 tratará das estratégias e técnicas de *prompt* escolhidas para a instrução no LLM a fim de harmonizar os termos e conceitos.

### 5.1 Linguística de *Corpus*

A Linguística de *Corpus* (LC) é uma das abordagens metodológicas mais utilizadas na área da Linguística, sobretudo na Linguística Computacional, para se estudar e analisar os fenômenos da língua por meio de *softwares* tanto de forma quantitativa como qualitativa (McCarthy; O’Keeffe, 2010). Tal enfoque surge a partir do desenvolvimento e do avanço tecnológico, facilitando o trabalho de coleta e análise de grandes quantidades de textos e permitindo que o linguista investigue e compreenda mudanças e padrões sintáticos, morfológicos, semânticos, discursivos, dentre outras questões, em inúmeras linhas de concordância, listas de frequência de palavras e palavras situadas em um contexto de uso (Shepherd, 2012). McEnery e Hardie (2012) salientam que o desenvolvimento da Linguística de *Corpus*, com suas técnicas, seus métodos e seus procedimentos, facilitou a exploração e criação de novas teorias linguísticas que se baseiam na linguagem em uso.

Embora a Linguística de *Corpus*, atualmente, esteja ligada à tecnologia de ponta e a ferramentas computacionais cada vez mais robustas, as primeiras evidências de uso

de *corpus* em análises linguísticas datam aos estudiosos bíblicos e literários, no século XIII, que buscavam identificar, linha por linha e palavra por palavra, palavras e frases em diferentes contextos para indexação (McCarthy; O’Keeffe, 2010). Posteriormente, na década de 50, os recursos da LC foram utilizados por estruturalistas americanos para compilar e analisar dados reais e em uso em suas pesquisas. De acordo com Bonelli (2010), com a chegada da Internet e com a transformação digital, a Linguística de *Corpus* se remodelou e se estabeleceu na área da Linguística. Em consonância a isso, Stefanowitsch (2020) ressalta que, mesmo que os estudos e os procedimentos com *corpus* nunca tenham cessado, a abordagem baseada na LC voltou a ocupar um lugar mais central na Linguística há pouco tempo.

Conforme explicita Reppen (2010), utilizar um *corpus* em trabalhos linguísticos é fundamental para investigar questões relativas ao uso da linguagem. Nesse sentido, o *corpus* permite que se analisem fenômenos linguísticos em contexto natural e de uso que não seriam perceptíveis com outro tipo de recurso (Mendes, 2016; McEnery; Hardie, 2012). Assim, a Linguística de *Corpus* pode ser definida como um método científico para analisar as nuances da língua. Para Brezina (2018, p. 6), o

*Corpus* (pl. *corpora*) é uma forma específica de dado linguístico. É uma coleção de textos escritos ou transcrições de linguagem falada que pode ser pesquisada por um computador utilizando um *software* especializado. Um *corpus*, geralmente, representa uma amostra de linguagem, i.e. um (pequeno) subconjunto de produção de linguagem de interesse, em alguns casos limitados de *corpora* muito especializados.

Para a compilação de um *corpus*, Hunston (2002) elucida que ele pode ser definido em termos de forma e propósito, isto é, a depender da finalidade da pesquisa, delineia-se um *corpus* que satisfaça aos requisitos e aos objetivos da investigação. Logo, um *corpus* é planejado e projetado para algum propósito linguístico. Pensando nisso, a construção de um *corpus* deve considerar alguns aspectos fundamentais: (i) a modalidade – escrito ou oral –; (ii) as línguas que o compõem e a sua classificação – monolíngue, bilíngue ou multilíngue –; (iii) o tipo – paralelo ou comparável; (iv) os gêneros a serem abarcados; e (v) a especificidade – de língua geral ou especializado. A definição desses pontos é crucial para a compilação e, posteriormente, para a obtenção de dados condizentes com a pesquisa a ser feita.

No que diz respeito à modalidade, um *corpus* pode ser composto por textos escritos ou transcrições de registros orais em formato eletrônico (Mendes, 2016). Com

relação ao segundo ponto, ao definir a língua ou as línguas do *corpus*, estaremos, também, classificando-o. Nesse sentido, um *corpus* composto por apenas uma língua é chamado de monolíngue; quando composto por um par de línguas, é denominado de bilíngue (geralmente utilizado em trabalhos de tradução), e um *corpus* composto por mais de duas línguas é chamado de multilíngue.

No que se refere ao tipo de *corpus*, podemos encontrar um *corpus* paralelo, ou seja, um conjunto de textos composto por textos originais em uma língua e a sua tradução na língua-alvo, ou um *corpus* comparável, que contém textos em uma língua e textos na outra língua. Destaca-se, aqui, que um *corpus* comparável não é composto por traduções, mas sim por textos do mesmo gênero, que não são idênticos. O quarto ponto está ligado aos gêneros que compõem o *corpus*. Assim, selecionam-se gêneros textuais que se relacionam com a pesquisa e que forneçam dados para a análise.

Acerca da especificidade do *corpus*, define-se se ele será de língua geral ou específico. Um *corpus* de língua geral deverá conter diferentes tipos de gêneros textuais a fim de representar diferentes variações da língua. Um exemplo de *corpus* de língua geral é o COCA<sup>39</sup> (*Corpus of Contemporary American English*), composto por mais de um bilhão de palavras e que busca retratar o inglês americano contemporâneo. Por outro lado, um *corpus* especializado não será tão extenso como um *corpus* de língua geral, uma vez que se considera o objetivo da pesquisa para a sua elaboração. Nesse sentido, um *corpus* especializado de um domínio deverá conter textos que sejam representativos para aquela área. De acordo com Koester (2010), esse tipo de *corpus* apresenta mais diretamente o contexto em que os textos foram produzidos e revela informações mais específicas como o cenário, os participantes, o propósito da comunicação e, principalmente, a possibilidade de indicar as relações entre os padrões linguísticos e os contextos de uso. Vale ressaltar que a definição da especificidade também está diretamente ligada ao quarto ponto, uma vez que um *corpus* mais geral necessita de diferentes tipos de gêneros textuais para dar conta de todos os padrões e as variações linguísticas, enquanto, em um *corpus* especializado, o gênero textual será mais focado na área a ser representada.

Além de definir esses pontos, é importante refletir sobre o tamanho, a representatividade e o balanceamento do *corpus*. Tais aspectos estão intrinsecamente relacionados, já que um conjunto de textos pequeno pode não representar toda a

---

<sup>39</sup>Davis, Mark. The Corpus of Contemporary American English (COCA). 2008. Disponível em: <https://www.english-corpora.org/coca/>. Acesso em: 20 maio 2024.



variedade da língua, os padrões linguísticos e a riqueza terminológica. Em contrapartida, um *corpus* mais extenso e robusto pode apresentar muitas generalizações e não fornecer os dados específicos e necessários para a pesquisa. Nesse sentido, muito se discute sobre a representatividade. Conforme afirma Mendes (2016, p. 227), a compilação do *corpus* “tem sempre subjacente o princípio de que o *corpus* irá fornecer dados que permitem algum grau de generalização sobre uma língua ou sobre algum aspecto de variação de uma língua, isto é, pressupõe algum princípio de representatividade.”. Assim, para que o *corpus* seja representativo de uma língua, é necessário considerar as variedades e manifestações linguageiras produzidas por uma população usuária ou de um tipo de discurso (Ädel, 2020).

Um ponto interessante de se utilizar a Linguística de *Corpus* como abordagem metodológica é a sua natureza interdisciplinar, posto que seus procedimentos são usados por pesquisadores de diferentes áreas. Na Terminologia, a aproximação entre as pesquisas terminológicas e a Linguística de *Corpus* começou no final dos anos 80 e no início dos anos 90, com o surgimento de novas correntes da Terminologia, que consideravam o contexto, o uso da linguagem em textos especializados e os aspectos comunicativos e cognitivos em sua base (Bevilacqua, 2013). Com o advento da tecnologia, a Linguística de *Corpus* passa a oferecer diferentes formas de se explorar e analisar os termos em contexto por meio de *softwares* que permitem extrair a terminologia e analisar cada ocorrência em seu real uso.

Inicialmente, o uso de *corpus* na área terminológica ocorreu de forma lenta, sendo realizada, sobretudo, por empresas privadas e por grupos de pesquisa acadêmica. Do ponto de vista de Maciel (2013), no contexto acadêmico, a aproximação Terminologia-Linguística de *Corpus* busca explorar teorias, confirmar hipóteses, testar metodologias tanto na área da Linguística como na da Computação. Além da parte mais teórica e acadêmica, na prática, a Terminologia também utiliza a LC para a elaboração, por exemplo, de dicionários, glossários e manuais. Para exemplificar essa interdisciplinaridade, o projeto Termisul<sup>40</sup> explora os estudos terminológicos, tradutórios, lexicográficos e computacionais, criando recursos de terminologia – como

---

<sup>40</sup> O projeto Termisul está disponível em: <https://www.ufrgs.br/termisul/index.php> . Acesso em: 2 mar. 2024.

as bases de dados Terminológicas<sup>41</sup> – que utilizam a Linguística de *Corpus* como metodologia para a sua elaboração.

Além dessa relação com a Terminologia, a Linguística de *Corpus* também apresenta interface com a Ciência da Computação, como na Engenharia do Conhecimento e na Inteligência Artificial (Maciel, 2013). Segundo Oliveira e Müller (2013), a interação entre LC e Computação favoreceu o aumento de pesquisas terminológicas na área da Linguística Computacional e no desenvolvimento de ferramentas computacionais no campo de PLN, que utiliza diversos recursos da Linguística de *Corpus* para a criação de analisadores sintáticos e morfológicos, corretores ortográficos, tradutores automáticos, mineração de dados, dentre outras possibilidades.

A seguir, trataremos do planejamento e da compilação dos *corpora* para este trabalho.

### 5.1.1 Planejamento e compilação dos *corpora*

Tendo em vista o nosso objetivo, nesta seção, apresentaremos as definições relativas ao planejamento e à compilação dos *corpora* desta pesquisa. Em um primeiro momento, planejamos elaborar um *corpus* composto por materiais utilizados nas diferentes áreas do programa, como documentos de marketing, finanças, compra e vendas, recursos humanos, logística, pesquisa e desenvolvimento, contabilidade, dentre outras. Conforme afirma Mautner (2021), os documentos da área empresarial fazem parte do nosso dia a dia, seja na posição de prestador ou usuário de serviços, vendedor, empregado, funcionário ou comprador. Estamos constantemente em contato com a comunicação empresarial e conhecemos os mais diferentes gêneros que compõem essa área. Com a alta produção diária desses documentos em transações e negócios, acredita-se que o acesso a eles seja fácil, ainda mais com a quantidade de sites de empresas. Logo, é possível compilar um *corpus* robusto e representativo da área.

Ao planejarmos coletar tais documentos do âmbito empresarial para a compilação do nosso *corpus*, encontramos alguns percalços durante a busca desse material, sendo o principal deles a questão de confidencialidade. De acordo com

---

<sup>41</sup> Uma das bases de dados terminológicas elaboradas pelo Termisul é a “Combinatórias Léxicas da Gestão Ambiental”, disponível em: [https://www.ufrgs.br/termisul/cles\\_gestamb/index.php](https://www.ufrgs.br/termisul/cles_gestamb/index.php). Acesso em: 2 mar. 2024.

Mautner (2021, p. 328), “[...] no mundo corporativo, os riscos financeiros e legais costumam ser maiores e, portanto, é mais difícil obter acesso, registrar, processar, armazenar e publicar dados que originalmente não deveriam estar na esfera pública”. Assim, a tarefa de compilar um *corpus* extenso, que abranja as diferentes áreas de uma empresa e com diferentes gêneros textuais é inexecutável. Dentre os documentos disponíveis *online* por empresas, encontramos relatórios anuais, relatórios de responsabilidade social corporativa, chamada de lucros, carta aos acionistas, relatórios de analistas financeiros, folhetos corporativos, relatórios para investidores e livros didáticos sobre negócios.

Considerando isso, para este trabalho, compilaremos dois *corpora* especializados, um voltado para a descrição do ERP (*Enterprise Resource Planning*), e o outro, *Corpus* para Harmonização, focado na linguagem da área de negócios. Com relação ao *Corpus* do ERP, coletamos manuais de instrução e materiais didáticos sobre o *software* para que possamos compreender o domínio, mapear os *frames* semânticos e construir a arquitetura de *Knowledge Graph* com *frames*. O *Corpus* para Harmonização, voltado para a linguagem da área de negócios, é composto por dicionários, glossários especializados de tal campo e documentos empresariais disponíveis *online* para que possamos analisar a frequência das palavras, extrair os termos, identificar as relações entre eles e selecionar a terminologia a ser harmonizada. Nesse ponto, também utilizaremos os termos específicos coletados do *Corpus* do ERP, visto que são unidades fundamentais para “popular” os *frames* e o KG com informações lexicais. Além disso, para a análise dos termos a serem harmonizados, também utilizaremos as informações coletadas da base de termos da empresa para buscarmos as traduções e as definições terminológicas. Desse modo, nossos *corpora* foram planejados conforme registra o Quadro 4.

Quadro 4 - Descrição dos *corpora*

	Modalidade	Gênero textual	Língua
<i>Corpus</i> do ERP	Escrito	Manuais de instrução do ERP; Materiais didáticos sobre o ERP.	Inglês
	Escrito	Dicionários especializados da área de negócios; Guias para gerenciamento de empresas; Glossários de termos do ERP.	Inglês

Corpus para Harmonização	Escrito	Relatório anual de empresas, comunicados de imprensa ( <i>press release</i> ), relatório financeiro, transcrições de conferências e reuniões e análise de resultados.	Inglês
--------------------------	---------	---	--------

Fonte: Elaborado pela autora.

Nossos *corpora* são específicos da área de negócios e dos departamentos abrangidos pelo ERP. Além disso, os *corpora* são em inglês e de modalidade escrita. De acordo com Koester (2010), alguns pontos podem tornar um *corpus* especializado, como, por exemplo, o objetivo específico da compilação, a contextualização, o gênero, o tipo de discurso, o assunto e a variedade da língua. No nosso caso, os *corpora* são especializados devido ao nosso objetivo de harmonizar termos e conceitos das áreas que compõem o ERP, bem como a contextualização da pesquisa, os gêneros selecionados do mundo corporativo, o tipo de discurso dos textos e o assunto a ser tratado nos documentos.

No que se refere aos gêneros dos *corpora*, salientamos que não foi possível obter acesso aos diversos documentos empresariais. Inicialmente, o planejamento do *Corpus* do ERP contava com os mais diversos gêneros produzidos nos setores que compõem o *software*. No entanto, não foi possível encontrar tais documentos, uma vez que eles são confidenciais. Como evidenciado por Mautner (2021), um dos grandes obstáculos de lidar com o discurso empresarial é a confidencialidade envolvida, ainda mais pelo risco de vazamento de informações financeiras, legais e estratégias de negócio. Justamente por isso, coletamos apenas documentos de empresas disponíveis na área para investidores, em virtude de que essas informações são de domínio público e são essenciais para realizar investimentos na empresa. É importante ressaltar também que todos os documentos, livros, dicionários e manuais estão disponíveis na Internet e podem ser acessados livremente.

A questão dos gêneros que compõem o *corpus* também se reflete na representatividade. Geralmente, pretende-se compilar um conjunto de textos representativo da área, buscando variados gêneros de diferentes fontes. Por se tratar de um *corpus* especializado e tendo em vista o objetivo deste trabalho, o tamanho e a quantidade de textos devem ser suficientes para dar conta do estudo (McEnery; Hardie, 2012). Mais além, Koester (2010) afirma que, em um *corpus* especializado, devemos levar em conta o grau de variabilidade acerca dos gêneros que o compõem, garantindo que os diversos documentos sejam incluídos na coleta. A título de exemplo, pode-se

coletar relatórios anuais de apenas uma empresa e construir um *corpus* desse gênero, ou, então, compilar um *corpus* de relatórios anuais de diversas empresas. No primeiro caso, o *corpus* pode ser representativo apenas do gênero utilizado em determinada empresa, mas ele não será a representação do gênero como um todo. No segundo caso, o *corpus* é representativo, visto que ele foi coletado de diferentes fontes e, assim, poderá evidenciar a variedade linguística do gênero. Para ilustrar uma parte da composição do *Corpus* para Harmonização, no Quadro 5, podemos verificar quais foram as empresas selecionadas para a coleta dos documentos e quais materiais foram obtidos.

Quadro 5 - Lista de empresas e documentos selecionados para compor o *Corpus* da harmonização

Empresas	Tipos de documento coletados
3M	Annual Financial Report
Adidas	Annual Report
Allbirds	Audit Committee Charter
Arc Document Solutions	Business and Strategic Updates
Audi	Business Partner Code of Conduct
BASF	Code of Conduct
Bed, Bath & Beyond	Code of Ethics
Best Buy	Compensation Report
BMW	Compliance Package
Bosch	Conference Transcriptions
Bradesco	Corporate Governance Report
Callaway	Corporate Responsibility Report
Carrefour	Earnings Presentation
Cirque du Soleil	Earnings Release
Coca-Cola	Earnings Release
Colgate	Factsheet
CVS Health	Financial Disclosures Reports
Foodstuffs North Island	Financial Statements
Gap Inc	Handling Unit Management for Supplier
Grupo Boticário	Investor Factbook
Harrolds	Investor Presentation
Johnson & Johnson	Inward Goods Carrier Information Pack (for Supply Chain Operations)
Liverpool	Management Report
Macy's Inc	Packaging Instructions
Mc Donald's	Press Release
Mercedes-Benz	Proxy Statement
Movida	Reports Fiscal Results
Natura	Sales Announcement
Neiman Marcus Group	Shareholder Letter
Nestlé	Social and Governance Report
Nike	Standards of Vendor Engagement
Nordstrom	Supplier Code of Conduct
Salling Group	Sustainability Reports
Sherwin-Williams	Tax Strategy Statements
Siemens	Terms of Implementation for Business Partners
Starbucks	Textile Supplier Manual
Target	

Toyota Tractor Supply Company Virgin Enterprises Limited Walgreens Walmart Wendy's Woolworths Group Zalando	Vendor Guide Web User Manual
--	---------------------------------

Fonte: Elaborado pela autora.

A partir do Quadro 5, podemos notar que nosso *Corpus* para Harmonização, por exemplo, será, em certa medida, representativo no que se refere aos documentos coletados de empresas. Ainda que não tenhamos coletado todos os documentos que fazem parte do âmbito empresarial, conseguimos encontrar uma grande diversidade de documentos, que podem evidenciar fenômenos linguísticos pertinentes para a harmonização de termos. Além disso, como discutido anteriormente, nossa coleta não se concentrou em apenas uma empresa, visto que coletamos diferentes documentos de muitas corporações.

Outro ponto que diz respeito ao *corpus* especializado é a vantagem de poder investigar os fenômenos de forma contextualizada, apontando para a estreita relação entre linguagem e contexto. De acordo com McCarthy e O'Keefe (2010), um *corpus* desse estilo apresenta os padrões linguísticos investigados atrelados ao contexto social e cultural de onde foram retirados. Nesse sentido, para o trabalho de propor *frames*, bem como para a identificação terminológica, um *corpus* especializado contribui para compreender o domínio em que as unidades lexicais estão inseridas e suas relações de forma contextualizada (L'Homme, 2020).

Com as discussões sobre a Linguística de *Corpus* e o planejamento e a compilação dos *corpora* desta pesquisa, na próxima seção apresentaremos as ferramentas utilizadas para o processamento e análise dos textos.

### 5.1.2 Ferramentas para extração de terminologia e análise de termos

Após o planejamento e a compilação dos *corpora*, nossa próxima etapa consiste em utilizar as ferramentas computacionais Sketch Engine<sup>42</sup> e LancsBox<sup>43</sup> para o

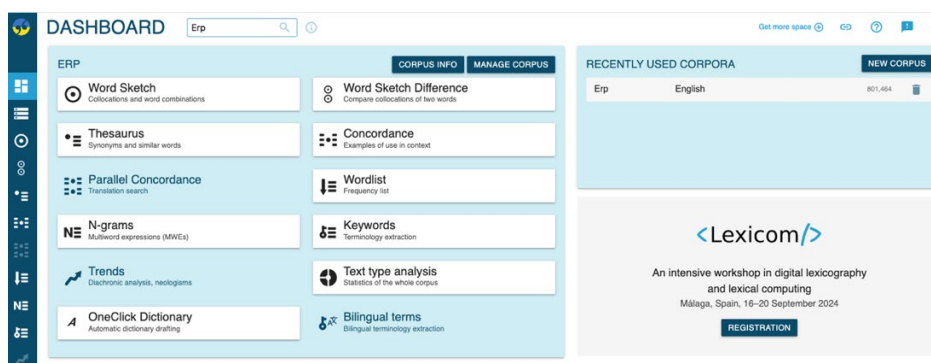
<sup>42</sup> A ferramenta Sketch Engine está disponível para compra em: <https://www.sketchengine.eu/>. Acessado em: 2 mar. 2024.

<sup>43</sup> A ferramenta LancsBox está disponível gratuitamente em: <http://corpora.lancs.ac.uk/lancsbox/>. Acesso em: 2 mar. 2024.

processamento e a análise dos dados. Optamos por utilizar tais *softwares*, uma vez que cada um apresenta diferentes recursos de análise.

A ferramenta Sketch Engine (Kilgarriff *et al.*, 2014) é um recurso pago que permite explorar como a língua funciona por meio da análise de grandes quantidades de texto (*corpus*), identificando padrões, gerando listas de palavras, extraindo termos, apresentando concordanciadores, dentre outras possibilidades. Esse *software* é utilizado por linguistas, lexicógrafos, tradutores, estudantes e professores justamente pela facilidade em se analisarem aspectos linguísticos que, sem o auxílio computacional, seriam difíceis de se observar. Na Figura 32, podemos visualizar o painel do Sketch Engine.

Figura 32 - *Dashboard* da ferramenta Sketch Engine



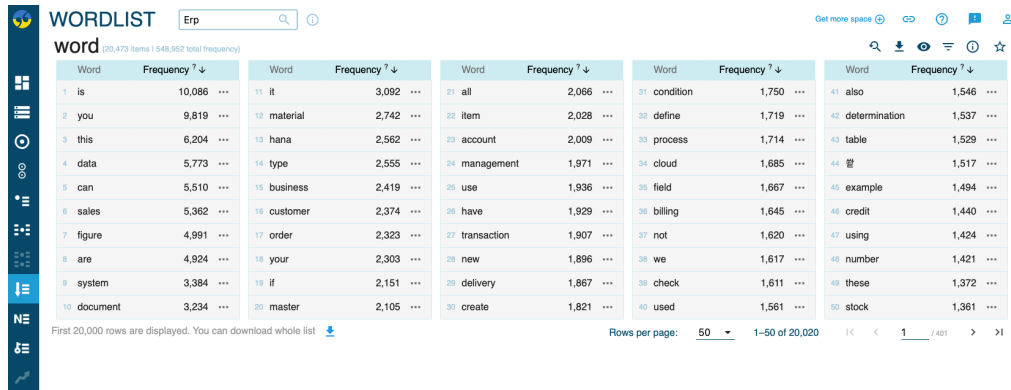
Fonte: Kilgarriff *et al.* (2014)<sup>44</sup>.

No painel da ferramenta, encontramos as seguintes opções de análise dos dados: *Word Sketch*, que explora as colocações e as combinações de palavras do *corpus*; *Word Sketch Difference*, que verifica e compara as colocações de duas palavras; *Thesaurus*, que apresenta os sinônimos e as palavras similares; *Concordance*, o concordanciador da ferramenta, o qual dispõe os exemplos de uso em contexto; *Parallel Concordance*, utilizado para pesquisas tradutórias; *Wordlist*, que apresenta a lista de frequências da palavras do *corpus*; *N-grams*, que explora as expressões com mais de uma palavra (*multi-word expressions*); *Keywords*, que extrai a terminologia do *corpus*; *Trends*, que trata da análise diacrônica e de neologismos; *Text type analysis*, que apresenta uma análise estatística do *corpus*; *OneClick Dictionary*, que produz um dicionário automaticamente com os dados do *corpus*; e, por fim, *Bilingual terms*, que extrai a terminologia bilíngue do *corpus*.

<sup>44</sup>LEXICAL COMPUTING. Sketch Engine. República Tcheca, 2014.

Para este trabalho, utilizaremos a ferramenta Sketch Engine para a extração da lista de frequência das palavras dos *corpora*, *Wordlist*, conforme a Figura 33.

Figura 33 - *Wordlist* do *corpus* gerada pelo Sketch Engine



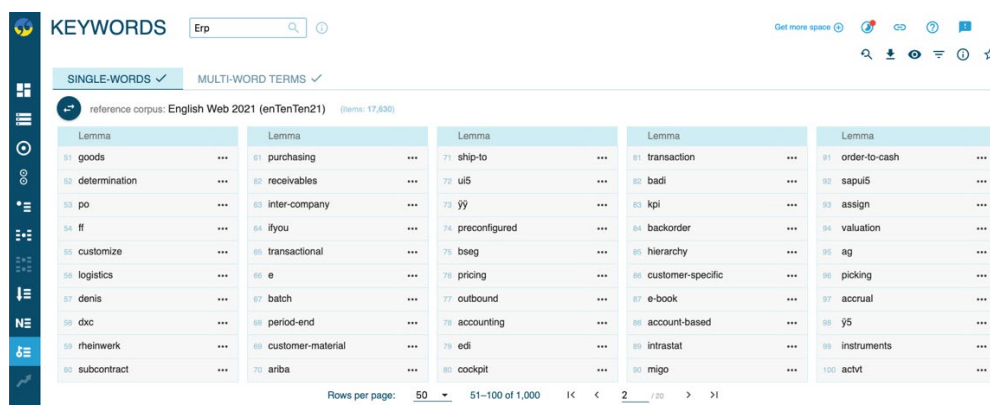
The screenshot shows the 'WORDLIST' interface for the 'Erp' corpus. It displays a table with 20,020 rows of words and their frequencies. The table is organized into five columns, each with a 'Word' and 'Frequency' header. The first column shows the top 10 words, including 'is' (10,086), 'you' (9,819), 'this' (6,204), 'data' (5,773), 'can' (5,510), 'sales' (5,362), 'figure' (4,991), 'are' (4,924), 'system' (3,384), and 'document' (3,234). The second column shows words like 'it' (3,092), 'material' (2,742), 'hana' (2,562), 'type' (2,555), 'business' (2,419), 'customer' (2,374), 'order' (2,323), 'your' (2,303), 'if' (2,151), and 'master' (2,105). The third column includes 'all' (2,066), 'item' (2,028), 'account' (2,009), 'management' (1,971), 'use' (1,936), 'have' (1,929), 'transaction' (1,907), 'new' (1,896), 'delivery' (1,867), and 'create' (1,821). The fourth column lists 'condition' (1,750), 'define' (1,719), 'process' (1,714), 'cloud' (1,685), 'field' (1,667), 'billing' (1,645), 'not' (1,620), 'we' (1,617), 'check' (1,611), and 'used' (1,561). The fifth column shows 'also' (1,546), 'determination' (1,537), 'table' (1,529), 'example' (1,517), 'example' (1,494), 'credit' (1,440), 'using' (1,424), 'number' (1,421), 'these' (1,372), and 'stock' (1,361). The interface includes a search bar, a sidebar with navigation icons, and a footer indicating 'First 20,000 rows are displayed. You can download whole list' and 'Rows per page: 50'.

Fonte: Kilgarriff *et al.* (2014).

A partir da geração de listas de frequência, é possível observarmos quais são as palavras mais frequentes e selecionarmos quais serão analisadas na pesquisa, visto que a ferramenta permite que se gerem listas de frequência de diferentes classes, como, por exemplo, substantivos, verbos, adjetivos e outras classes gramaticais. Para a tarefa de harmonização, essas possibilidades são essenciais para podermos selecionar os termos, analisarmos quais são os mais recorrentes e classificarmos de acordo com a sua classe gramatical.

Além das listas de frequência, é imprescindível recorrermos ao *Keywords*, conforme Figura 34, a fim de extrairmos a terminologia do *corpus* para, então, analisarmos e selecionarmos os termos a serem harmonizados.

Figura 34 – Lista de *keywords* gerada pela ferramenta Sketch Engine



The screenshot shows the 'KEYWORDS' interface for the 'Erp' corpus. It displays a list of lemmas under the 'SINGLE-WORDS' tab. The interface includes a search bar, a sidebar with navigation icons, and a footer indicating 'reference corpus: English Web 2021 (enTenTen21) (Items: 17,636)' and 'Rows per page: 50'. The list of lemmas is organized into five columns, each with a 'Lemma' header. The first column shows 'goods', 'determination', 'po', 'ff', 'customize', 'logistics', 'denis', 'dxc', 'rheinwerk', and 'subcontract'. The second column shows 'purchasing', 'receivables', 'inter-company', 'ifyou', 'transactional', 'e', 'batch', 'period-end', 'customer-material', and 'ariba'. The third column shows 'ship-to', 'ui5', 'y', 'preconfigured', 'bseg', 'pricing', 'outbound', 'accounting', 'edi', and 'cockpit'. The fourth column shows 'transaction', 'badi', 'kpi', 'backorder', 'hierarchy', 'customer-specific', 'e-book', 'account-based', 'intrastat', and 'migo'. The fifth column shows 'order-to-cash', 'sapui5', 'assign', 'valuation', 'ag', 'picking', 'accrual', 'y5', 'instruments', and 'actvt'.

Fonte: Kilgarriff *et al.* (2024).



Outro recurso que utilizamos foi o *Word Sketch*, representado pela Figura 35, para gerarmos as colocações e verificarmos o seu comportamento nos textos. Isso nos permite compreender em quais situações e contextos uma determinada colocação está inserida e quais são suas relações com outras palavras do *corpus*.

Figura 35 - *Word Sketch* gerado pelo Sketch Engine

The screenshot shows the Word Sketch interface for the word "employee" (noun, 72x). The interface is divided into several columns, each representing a different grammatical category. The categories and their associated words are as follows:

Category	Words
modifiers of "employee"	full-time, responsible, tech-savvy, frustrated, interaction, qualified, behavior, ofyour, strong, contact, supplier, center
nouns modified by "employee"	turnover, satisfaction, investment, base, supplier, benefit, vendor, hierarchy, role, release, master, integration
verbs with "employee" as object	hire, denote, retain, engage, influence, establish, involve, need, manage, require, be
verbs with "employee" as subject	verify, purchase, receive, work, appear, maintain, check, have, be
"employee" and/or ...	supplier, courier, FTE, interaction, website, position, close, customer, allocation, credit, good
prepositional phrases	... by "employee", ... of "employee", "employee" of ..., ... to "employee", ... for "employee", "employee" with ..., ... with "employee", ... as "employee", "employee" on ..., "employee" by ..., ... from "employee", ... about "employee"
adjective predicates of "employee"	responsible

Fonte: Kilgarriff *et al.* (2024).

Por fim, utilizaremos o recurso Thesaurus para examinarmos as palavras similares e mais próximas, conforme apresenta a Figura 36.

Figura 36 - Recurso Thesaurus da ferramenta Sketch Engine

The screenshot shows the Thesaurus interface for the word "employee" (noun, 72x). It displays a list of related words (Lempos) and their frequencies. The words are arranged in a grid format, with columns for Lempos and Frequency. The words listed are:

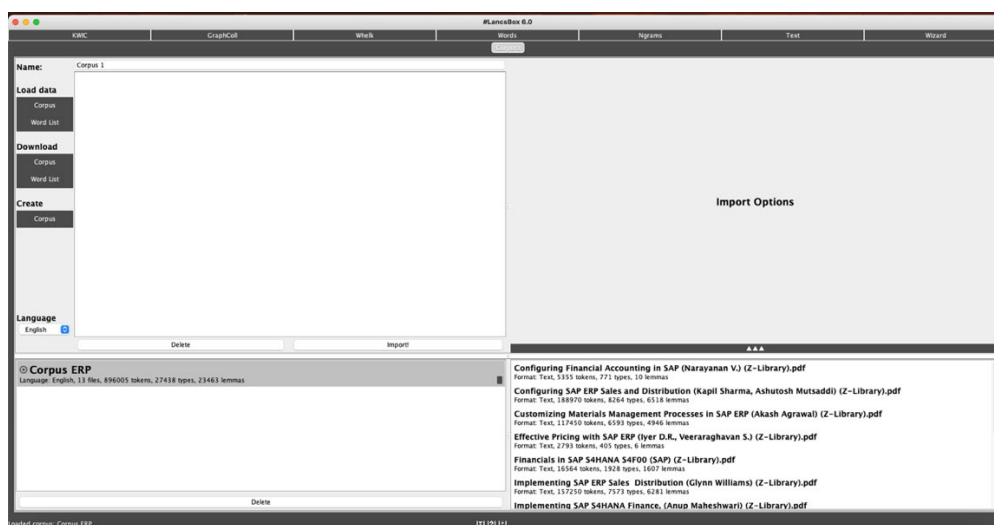
Lempos	Frequency	Lempos	Frequency	Lempos	Frequency	Lempos	Frequency	Lempos	Frequency
1 person	79	11 Materials	305	21 namespace	16	31 Code-2	14	41 accountant	60
2 buyer	26	12 satisfaction	9	32 USD	97	32 communication	112	42 country	188
3 runtime	35	13 liquidity	178	33 partner	999	33 hierarchy	522	43 backend	35
4 agent	24	14 specialist	14	34 purchase	517	34 RL00	9	44 Purchasing	277
5 individual	23	15 ship-to	29	35 decision-making	31	35 non-SAP	92	45 vendor	985
6 bank	374	16 adaptation	14	36 Sharma	9	36 PRO00	21	46 download	17
7 GL	31	17 info	156	37 NIMBL	34	37 premise	29	47 addition	173
8 people	65	18 analytics	42	38 security	75	38 address	123	48 house	65
9 clerk	12	19 team	177	39 master	1,567	39 site	32	49 Ifyou	22
10 department	164	20 Post	58	40 inquiry	79	40 district	16	50 XYZ	17

Fonte: Kilgarriff *et al.* (2024).

Além da ferramenta Sketch Engine, também utilizaremos a ferramenta LancsBox, desenvolvida pela Universidade de Lancaster. Tal *software*, disponível gratuitamente, pode ser usado por linguistas, professores de línguas, historiadores, sociólogos, educadores e demais pessoas interessadas em estudar e compreender os

aspectos da língua. O recurso possibilita que visualizemos os dados linguísticos e os analisemos. Na Figura 37, podemos observar o painel da ferramenta LancsBox.

Figura 37 - Ferramenta LancsBox

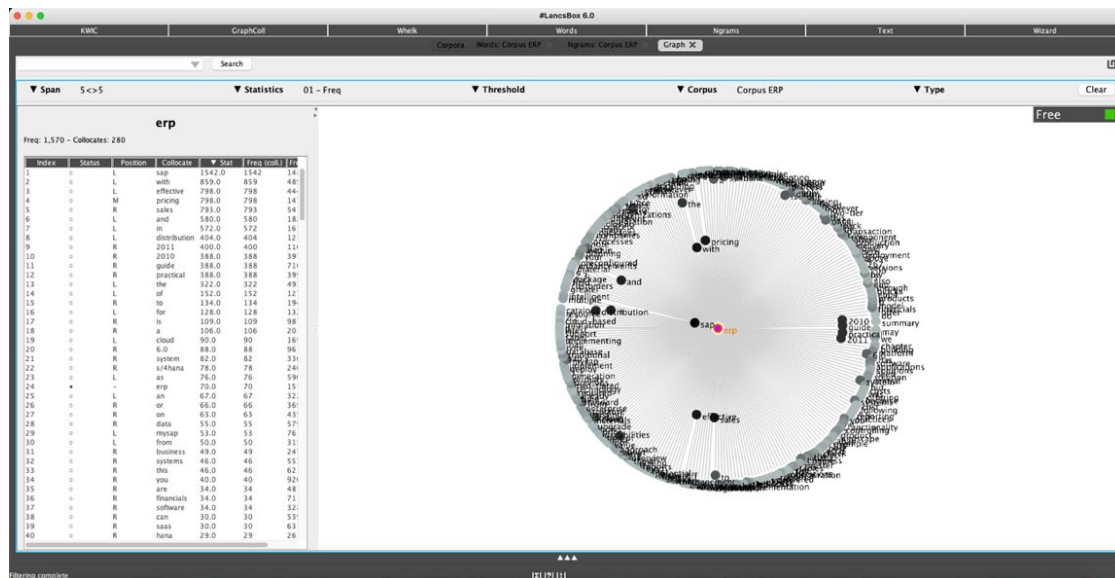


Fonte: Brezina, Weill-Tessier; McEnery (2020)<sup>45</sup>.

Esse programa computacional permite que se analisem os dados linguísticos por meio de diferentes opções, como, por exemplo, *KWIC*, que gera lista de todas as instâncias de um termo no formato de concordanciador; *Whelk*, que fornece informações sobre como um termo está distribuído no *corpus*; *Words*, que verifica os tipos de lemas e classes gramaticais no *corpus*; *GraphColl*, que identifica as colocações e apresenta as palavras mais próximas de um determinado termo pesquisado no formato de grafo; *Text*, que apresenta os exemplos de termos em contexto; *N-grams*, que explora as expressões compostas por mais de uma palavra; e, finalmente, *Wizard*, que combina todos os recursos da ferramenta e produz um relatório de pesquisa acerca do *corpus* analisado.

Para a nossa pesquisa, além dos recursos disponíveis da ferramenta Sketch Engine, também utilizaremos a opção *GraphColl* do LancsBox, visto que ela permite que se examinem as relações entre os termos e veja quais são as ocorrências mais próximas do termo pesquisado, conforme a Figura 38.

<sup>45</sup> BREZINA, V.; WELL-TESSIER, P.; MCENERY, A. #LancsBox v.5.x. Lancaster, 2020. *Software*.

Figura 38 - Recurso *GraphColl* gerado pela ferramenta LancsBox

Fonte: Brezina; Weill-Tessier; McEnery (2020)<sup>46</sup>.

Esse recurso não é oferecido na ferramenta Sketch Engine e é essencial para compreendermos o comportamento dos termos no *corpus*, bem como refletirmos sobre as relações entre eles. Para a nossa tarefa de harmonização e para a construção da arquitetura de *Knowledge Graph* e *frames* semânticos, analisar e compreender as relações entre os termos é crucial para podermos visualizar e selecionar os dados a serem trabalhados.

Após descrevermos as ferramentas computacionais utilizadas para o processamento e a análise dos *corpora*, assim como detalhar quais recursos serão empregados para a extração dos termos a serem harmonizados e para compreender o comportamento deles em um contexto de uso, a seguir, trataremos dos *Knowledge Graphs* e dos *Frames*, apresentando como serão utilizados no trabalho e como construiremos tal arquitetura.

## 5.2 *Knowledge Graphs* e *Frames* semânticos

Além de trabalharmos na parte teórica da harmonização de termos, nossa pesquisa também apresenta uma parte computacional aplicada, isto é, criaremos uma arquitetura de *Knowledge Graph* (KG) com *frames* semânticos, que será utilizada tanto na base de dados da empresa, de forma experimental, com o objetivo de estruturar a

<sup>46</sup> BREZINA, V.; WELL-TESSIER, P.; MCENERY, A. #LancsBox v.5.x. Lancaster, 2020. *Software*.

representação do conhecimento do ERP e para a recuperação da informação. Além disso, o KG será implementado em um LLM (*Large Language Model*), a fim de que a harmonização seja feita de forma automática. Essa etapa surge, principalmente, para que se trabalhe a harmonização e a sua implementação, extrapolando a proposta metodológica da ISO, ou seja, indo além de uma mera análise de termos e uma comparação entre eles, mas que também se considerem os aspectos contextuais e situacionais intrínsecos aos termos e que devem ser examinados na análise e na proposição da harmonização. Nesse sentido, o KG permitirá que criemos arquitetura processável e legível por máquina, visto que será implementada tanto na base de dados da empresa, experimentalmente, como no LLM, e os *frames* semânticos adicionarão uma camada extra de informação linguística, que, muitas vezes, não é contemplada na elaboração de um grafo.

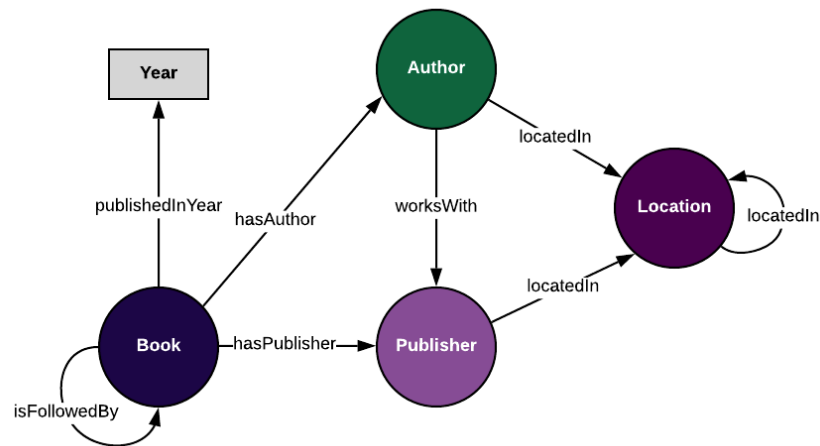
Antes de tratarmos da criação do *Knowledge Graph* com *frames*, vale discutirmos sobre a estrutura do KG. Como vimos no capítulo anterior, esse tipo de arquitetura é uma forma de representar o conhecimento e pode ser referido como grafo conceptual e ontologia (Steth; Gyrard; Padhee, 2019). Sob a mesma perspectiva, Kejriwal e Knoblock (2021) afirmam que um *Knowledge Graph* é uma ontologia. À vista disso, torna-se crucial visualizarmos as estruturas para compreendermos como tal representação será aplicada neste trabalho, visto que optamos por elaborar um KG com *frames*.

Conforme explicita Schrader<sup>47</sup> (2020), ontologias são modelos semânticos que definem os tipos de coisas que existem em um determinado domínio e suas propriedades. Além disso, elas são modelos de dados genéricos, ou seja, uma ontologia somente modela tipos genéricos de coisas que compartilham certas propriedades, não incluindo informações específicas de um domínio. Assim, uma ontologia foca em conceitos mais gerais, capturando as características que podem ser utilizadas para descrever objetos e coisas do mundo aplicáveis a todos. A título de exemplo, na Figura 39, podemos observar uma ontologia, que descreve as classes, os atributos e as relações de um autor e suas publicações.

---

<sup>47</sup> SCHRADER, Bess. What's the difference between an Ontology and a Knowledge Graph?. Enterprise Knowledge [S.l.], 2020. Disponível em: <https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/>. Acesso em: 2 mar. 2024.

Figura 39 - Ontologia sobre informação de autor, livros e publicações



Fonte: Schrader (2020)<sup>48</sup>.

Neste exemplo de ontologia, podemos visualizar as classes e relações criadas entre o autor (*author*) de um determinado livro (*book*), o local de publicação (*location*), quem publicou (*Publisher*), o título do livro e o ano de publicação (*year*). Como classes, temos: autor, local, editora e livro. Já as relações são: tem-é (*has*), localizado em (*located in*), publicado em (*published in*). Considerando isso, para a classe livro (*book*), teremos as seguintes propriedades e relações:

#### Quadro 6 - Descrição da ontologia

Um livro (*book*) tem um autor (*author*).

Um livro tem uma editora (*Publisher*).

Um livro é publicado em um ano (*year*).

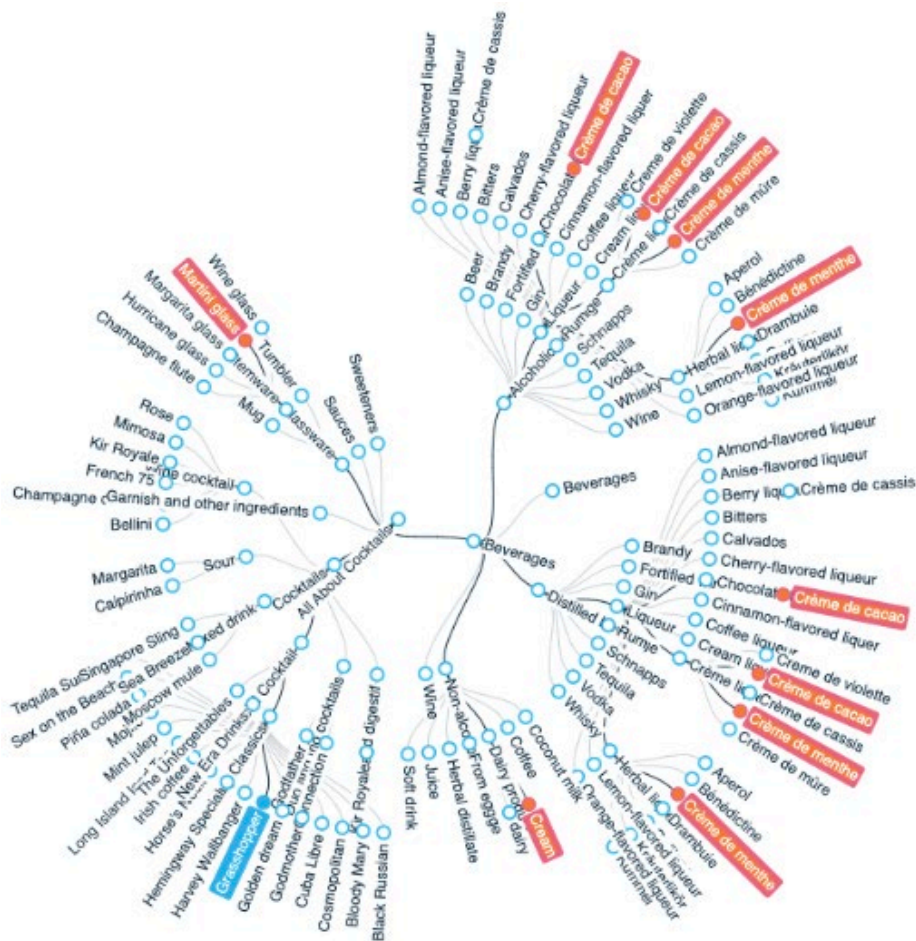
Um livro é seguido de outro livro (sequência).

Fonte: Elaborada pela autora.

A ontologia criada por Schrader (2020) mostra de forma simples como se estrutura tal representação do conhecimento, especialmente para indicar a questão de não se aprofundar em detalhes mais específicos. No exemplo da Figura 40, podemos visualizar uma ontologia mais robusta, que descreve os tipos de bebidas.

<sup>48</sup> SCHRADER, Bess. What's the difference between an Ontology and a Knowledge Graph?. Enterprise Knowledge [S.l.], 2020. Disponível em: <https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/>. Acesso em: 2 mar. 2024.

Figura 40 - Ontologia sobre bebidas

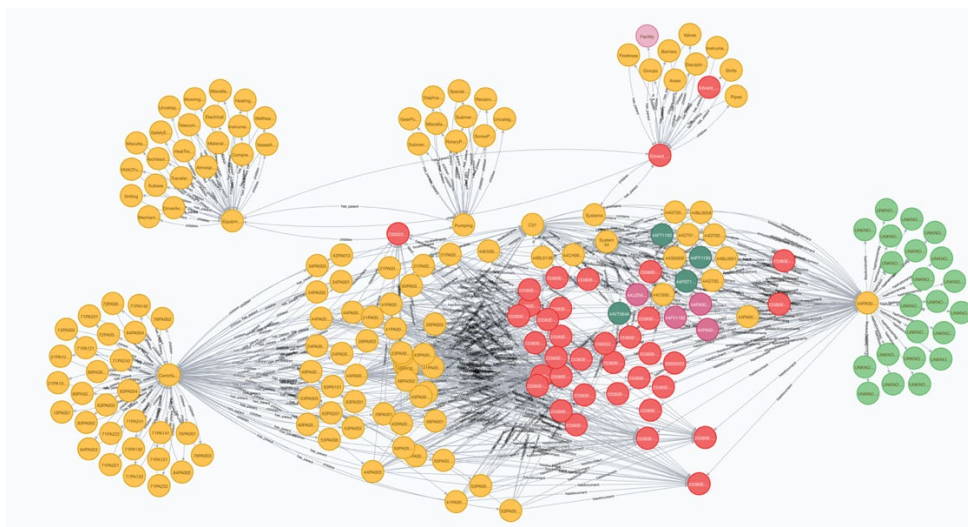


Fonte: Hedden (2023) <sup>49</sup>.

Nessa ontologia, podemos observar mais classes. No centro da arquitetura, temos a classe bebidas (*beverages*), com diferentes tipos: destilados (*distilled*), alcoólicos (*alcoholic*), não alcoólicos (*non-alcoholic*) e coquetéis (*All about Cocktails*). Em seguida, podemos observar as relações mais aprofundadas de cada classe. Ainda que a autora não tenha descrito as relações entre as classes, podemos ver que uma ontologia comporta uma grande quantidade de dados.

Como já discutido anteriormente, a estrutura de um *Knowledge Graph* é a mesma de uma ontologia, apresentando entidades e relações. O grande diferencial do KG é a possibilidade de conectar dados interligados e de diferentes bases. Para ilustrar essa conexão entre diferentes dados de KGs, a Figura 41 exemplifica a arquitetura de representação do conhecimento com diversos objetos interligados.

<sup>49</sup> HEDDEN, Heather. Taxonomies vs. Ontologies. Hedden Information Management. Carlisle, 2023. Disponível em: <https://www.hedden-information.com/taxonomies-vs-ontologies/>. Acesso em: 17 fev. 2024.

Figura 41 - Interligação entre *Knowledge Graph*

Fonte: *Knowledge Graph* [...] (2024) <sup>50</sup>.

Na Figura 42, podemos observar a grande quantidade de dados que um *Knowledge Graph* comporta, mas, principalmente, a interligação entre diferentes grafos. Para o nosso trabalho, essa possibilidade é fundamental, pois nos auxilia na construção do KG do ERP, conectando os diferentes setores que compõem o programa.

Após trazeremos aqui apresentação da estrutura de uma ontologia e de um *Knowledge Graph*, trataremos da metodologia para a criação do KG com *frames* semânticos.

### 5.2.1 Criação do *Knowledge Graph* e *frames semânticos* e Harmonização de termos e conceitos

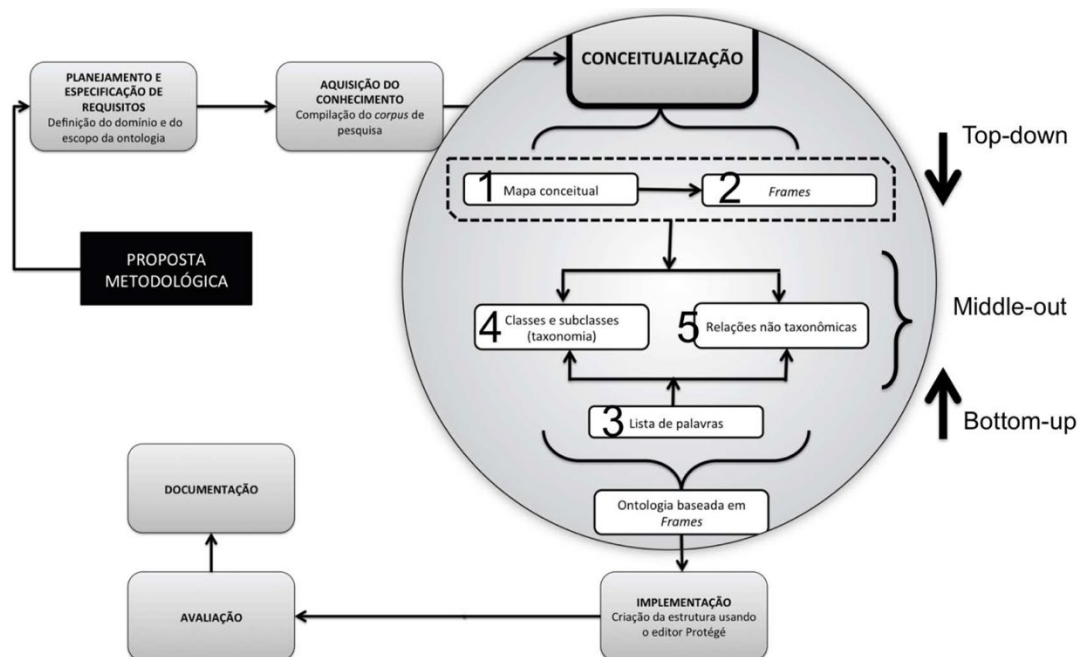
Como o nosso objetivo é construir um *Knowledge Graph* com *frames* semânticos para que seja implementado de forma experimental na base de dados da empresa, bem como para utilizarmos como estratégia de *prompt* no LLM, trataremos da metodologia de proposição dessa arquitetura. No capítulo anterior, quando discorreremos sobre representação do conhecimento, apresentamos brevemente a proposta metodológica para construção de uma ontologia jurídica baseada na Semântica de *Frames* de Müller (2015). Como já afirmamos, a elaboração do nosso KG e *frames*

<sup>50</sup> *KNOWLEDGE GRAPH*. In.: Eigen. [S.l.], 2024. Disponível em: <https://eigen.co/knowledge-graphs/>  
Acesso em: 17 fev. 2024.

fundamentar-se-á na proposta da autora. Logo, torna-se imprescindível aprofundarmos nas etapas criadas por Müller.

De acordo com Müller (2015), sua metodologia para a proposição de uma ontologia com *frames* consiste em cinco etapas: (i) planejamento e especificação dos requisitos, (ii) aquisição do conhecimento, (iii) conceitualização, (iv) implementação e (v) avaliação. Na Figura 42, podemos observar todo o processo de modelação da arquitetura.

Figura 42 - Etapas metodológicas propostas de Müller



Fonte: Müller (2015, p. 102).

A primeira etapa metodológica prevista no trabalho de Müller consiste em planejar e especificar os requisitos. Para isso, a autora se baseia nas seguintes perguntas: que domínio a ontologia pretende cobrir e para que a ontologia será utilizada?

Em nosso trabalho, nosso KG cobrirá o *software* ERP (*Enterprise Resource Planning*), incluindo todas as áreas que ele abarca, i.e., recursos humanos, logística, compra e vendas, pesquisa e desenvolvimento, finanças, contabilidade, manufatura, cadeia de suprimentos, *sourcing & procurement* e serviços. É interessante destacar que podemos compreender cada área do ERP como uma subárea do programa. Nesse sentido, ao considerarmos todas as especificidades das diferentes áreas do *software*, cada uma representará um KG, ainda que compartilhe muitos eventos, participantes e características semelhantes. Desse modo, percebemos que a arquitetura desse



*Knowledge Graph* deverá contar com uma rede de grafos, uma vez que o ERP conecta todas as áreas, e isso poderá ser visto tanto na sua estrutura como na sua implementação. É justamente por essa rede interligada que um *Knowledge Graph* é mais apropriado para esse tipo de trabalho, seja pelo compartilhamento de dados entre os setores do ERP (*Data Fabric*), seja pela possibilidade de representar o conhecimento de maneira mais detalhada (Martin; Szekeley; Allemang, 2021).

Com relação à segunda pergunta, nosso *Knowledge Graph* será utilizado para auxiliar na harmonização de termos e conceitos, sendo também implementado experimentalmente na base de dados da empresa para que auxilie na organização da base. Por fim, será usado como uma das estratégias de *prompt* no LLM para que seja possível harmonizar os conjuntos de termos automaticamente, obtendo resultados mais precisos e fiáveis.

Na etapa de aquisição do conhecimento, segundo estágio da proposta metodológica de Müller (2015), busca-se compreender melhor o domínio a ser representado. Para isso, a autora trabalha com dois *corpora*: um composto por documentos jurídico, como manuais, *corpus* de consulta, e outro *corpus* constituído de termos de audiência. Tendo em vista o nosso objetivo de pesquisa, trabalharemos com dois *corpora*: o *Corpus* do ERP, composto por livros, manuais e documentos sobre o programa; e o *Corpus* para Harmonização, gerado a partir da coleta de dicionários e glossários do ERP, bem como de relatórios anuais, *press release*, documentos para investidores, dentre outros arquivos corporativos. Com o primeiro *corpus*, *Corpus* do ERP, estudaremos o domínio e definiremos os *frames*. O *Corpus* para Harmonização, por sua vez, servirá para extrairmos as listas de termos e coletarmos os exemplos em uso. Além disso, recorreremos ao *Corpus* do ERP para trabalharmos com termos específicos do programa.

A terceira etapa da abordagem de Müller (2015) consiste na conceitualização. Como destaca a autora, essa é a principal etapa para a construção de uma ontologia baseada em *frames*. Para esse ponto, busca-se construir um mapa conceitual do domínio, elaborar os *frames*, gerar listas de palavras e estabelecer as classes, subclasses e as relações não taxonômicas. De modo geral, de acordo com Müller (2015), a metodologia segue um direcionamento *middle-out*, isto é, tal procedimento une as abordagens *top-down* e *bottom-up*. Nesse sentido, o direcionamento *top-down* consiste, em um primeiro momento, na proposição de *frames* e, considerando a análise dos *corpora* compilados, na identificação dos termos. Em um segundo momento, o

direcionamento *bottom-up* consiste na extração da terminologia dos *corpora* e na análise qualitativa delas com relação à evocação dos *frames* identificados.

Acerca do mapa conceitual, a autora identificou todos os eventos que ocorrem no procedimento sumaríssimo, descrevendo cada um e analisando-os separadamente para examinar os participantes envolvidos. Ainda nesse estágio, também se identificaram as relações entre eventos, verificando se havia uma sequência entre eles. Após a análise, Müller partiu para a elaboração do mapa conceitual e, concomitantemente, identificou os conceitos ilustrados no mapa para a ontologia.

No que se refere à elaboração dos *frames*, após a criação do mapa conceitual, a autora passou a analisar quais seriam os *frames* relacionados ao procedimento sumaríssimo. Para isso, a autora buscou auxílio no mapa conceitual para a listagem das cenas apresentadas através dos eventos. Em um primeiro momento, a pesquisadora retomou as descrições dos eventos que fundamentaram o mapa conceitual a fim de construir os *frames* por meio de uma definição (glosa). Tal definição deve conter uma visão geral da cena que o *frame* descreve, além de apresentar o envolvimento dos participantes. Nessa etapa, a autora seguiu os princípios da FrameNet.

Em um segundo momento, Müller (2015) verificou quais eram os participantes de cada evento, assim como as demais informações para delimitar os elementos de *frame*. O último passo foi identificar as unidades lexicais evocadoras dos *frames* a partir dos documentos do *corpus* de consulta. Um ponto a ser tratado diz respeito às relações entre os *frames*, visto que isso colabora para a elaboração da estrutura taxonômica da ontologia, indicando as classes e subclasses. Voltando para o nosso trabalho, nesta etapa metodológica, utilizaremos o *Corpus* do ERP para a construção do mapa conceitual e dos *frames*.

A terceira etapa da proposta de Müller acerca da conceitualização condiz com a lista de palavra. Para isso, a autora analisou o *corpus* com o auxílio de ferramentas computacionais para a identificação de palavras ou termos relativos ao domínio. Neste trabalho, utilizaremos os dados dos *corpora* da Harmonização e do ERP para gerarmos listas de palavras e extrairmos os termos para “popular” nosso mapa conceitual e *frames*.

Na próxima etapa, a autora buscou identificar no mapa conceitual e nos *frames* as classes e subclasses para organizar a estrutura hierárquica da ontologia. O último passo da conceitualização é estabelecer as relações não taxonômicas a partir da análise dos *frames*. Esse passo permitiu que a autora pudesse descrever as relações entre

eventos e participantes, indicando as conexões entre classes e subclasses. Inicialmente, Müller (2015) analisou os participantes e os papéis exercidos em cada *frame* para, então, estabelecer as relações.

Após a etapa de conceitualização, Müller parte para o estágio de implementação. Nesse momento, a autora usa a ferramenta *Protégé* para elaborar a ontologia, considerando os dados linguístico-conceituais. O último estágio da proposta da autora consiste em avaliar a ontologia. Para isso, verificou-se automaticamente a ontologia, buscando examinar se havia inconsistências e, em seguida, foi realizada uma verificação manual.

Para a aplicação da proposta de Müller (2015) em nosso trabalho, durante a etapa de conceitualização, adicionaremos mais um passo que se refere à harmonização terminológica. Como descrito pela autora, concomitante ao mapeamento dos *frames* e à construção do mapa conceitual, extraíram-se palavras e termos para “popular” sua arquitetura. Nesse momento, além de trabalharmos na extração de termos, também nos ocuparemos da aplicação da metodologia semântico-terminológica da harmonização de termos e conceitos. Esse estágio de harmonização durante a conceitualização é fundamental em nosso trabalho, uma vez que o ponto alto da nossa concepção de harmonização é a adição dos *frames*, que passam a apresentar o contexto em que os termos estão inseridos. Tendo em vista que a proposta de harmonização desenvolvida pela ISO não considera o aspecto contextual na sua aplicação, a inserção do contexto via *frames* semânticos na metodologia que iremos propor é crucial para compreendermos o comportamento da terminologia do domínio, suas relações e as ligações entre os conceitos.

No que tange à etapa de harmonização, após analisarmos os trabalhos que desenvolveram essa tarefa terminológica (cf. Lima (2016), Segunda (2017) e Lisboa (2021)), descritos na subseção 3.3.1 Harmonização de conceitos e termos: metodologia e prática, que seguem os preceitos da ISO 860:2007 / NBR 13790:1997 em suas aplicações de harmonização terminológica, seguiremos os passos para harmonizar os conjuntos de dados terminológicos. Como veremos no Quadro 7, considerando os trabalhos analisados, após a coleta, compilação e processamento do *corpus*, para executar a harmonização, seguiram-se as seguintes etapas indicadas no Quadro 7.

Quadro 7 - Etapas da aplicação de harmonização terminológica

ETAPA	DESDOBRAMENTOS
(i) Extração de listas de termos;	
(ii) Elaboração de listas de candidatos a termo;	
(iii) Validação terminológica;	
(iv) Comparação entre a terminologia proposta antes da harmonização e a nova terminologia extraída;	Durante essa etapa, também são analisados variantes, sinônimos, falsos sinônimos e termos em desuso.
(v) Criação de fichas terminológicas;	As fichas terminológicas, legíveis por máquina, são compostas pelas seguintes informações: a) Identificação do termo; b) Entrada; c) Classe gramatical; d) Abreviatura/sigla; e) Domínio; f) Definição; g) Fonte da definição; h) Contexto; i) Sinônimos; j) Variantes; k) Equivalentes.
(vi) Desenvolvimento de uma base de dados harmonizada.	A base de dados deverá considerar as seguintes informações para a sua proposição: a) Público-alvo; b) Proposta (objetivo) de criação.

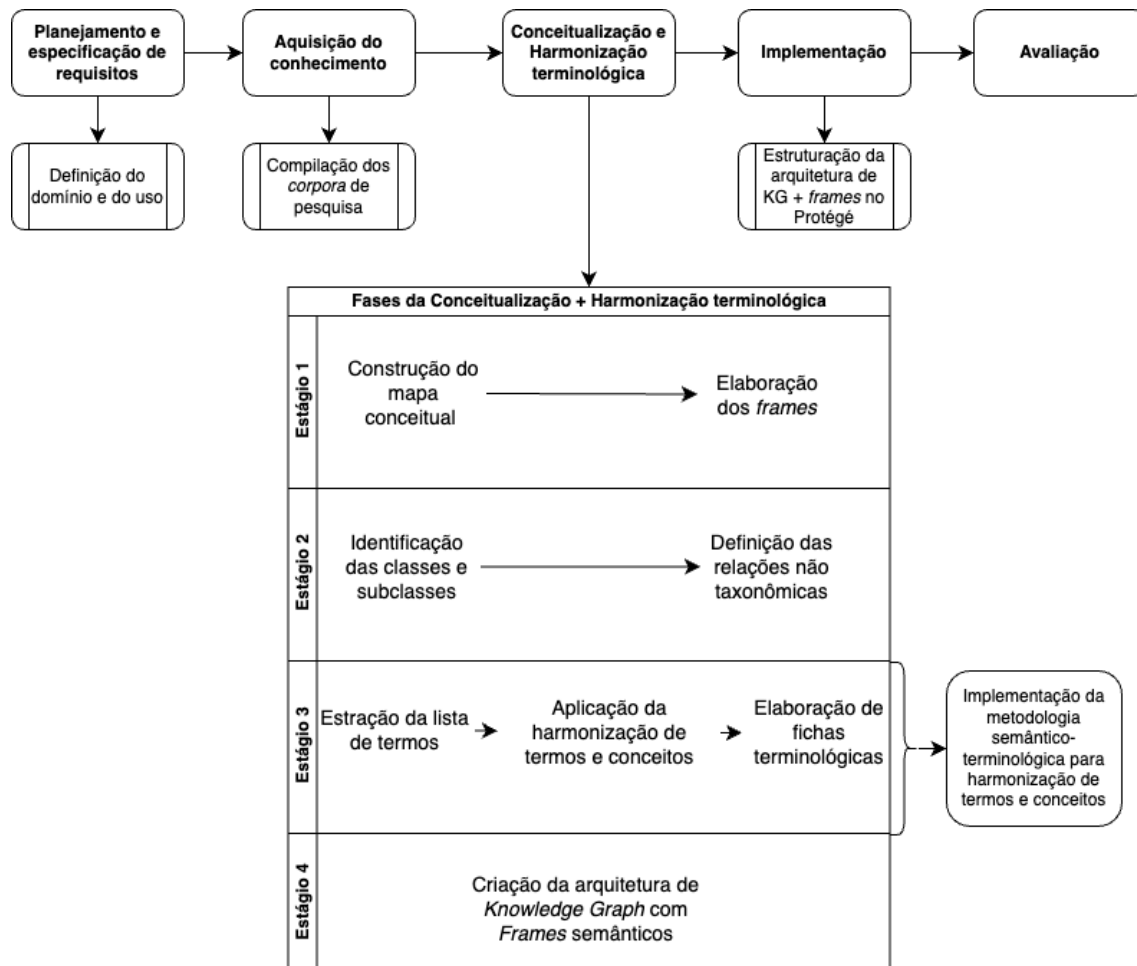
Fonte: Elaborado pela autora.

Considerando os passos seguidos nos estudos apresentados, para o desenvolvimento do nosso procedimento de aplicação de harmonização de termos e conceitos, planejamos as seguintes etapas:

- (i) Revisão do fluxograma proposto pela ISO 860:2007 / NBR 13790:1997 para aplicação da harmonização de termos e conceitos;
- (ii) Elaboração do mapa conceitual do ERP;
- (iii) Realização da descrição da arquitetura de *frames* semânticos;
- (iv) Extração da terminologia do ERP;
- (v) Análise preliminar da estrutura de *frames*;
- (vi) Validação da lista de termos;
- (vii) Comparação da terminologia desenvolvida antes do processo de harmonização com a nova listagem;
- (viii) Construção do *Knowledge Graph* com *frames*;
- (ix) Elaboração de fichas de cadastro dos termos.

Na imagem Figura 43, podemos observar as etapas metodológicas previstas para a construção do *Knowledge Graph* com *frames* semânticos, bem como as etapas de harmonização de termos.

Figura 43 - Etapas metodológicas para a construção do *Knowledge Graph* com *frames* e para a aplicação da harmonização terminológica



Fonte: Elaborada pela autora.

De acordo com a Figura 43, é possível examinar todas as etapas que seguiremos para a elaboração da arquitetura de representação do conhecimento e para a aplicação da harmonização de termos e conceitos. Partindo da proposta metodológica de Müller (2015), iniciamos com a etapa de planejamento e especificação de requisitos, definindo o domínio e o uso do KG. Em seguida, trabalhamos na aquisição do conhecimento, compilando os *corpora* da pesquisa. O terceiro passo é ocuparmo-nos da conceitualização. Essa etapa é vital tanto para a elaboração da arquitetura de representação do conhecimento como para a aplicação da metodologia semântico-terminológica da harmonização de termos e conceitos. A quarta etapa prevê a

implementação do *Knowledge Graph* com *frames* semânticos e, por fim, o quinto passo é avaliar a estrutura completa para verificar possíveis inconsistências e problemas.

A partir do detalhamento da metodologia criada por Müller (2015) para o desenvolvimento de uma ontologia baseada em *frames* e dos nossos passos metodológicos descritos até aqui, a seguir, apresentaremos as ferramentas computacionais que utilizaremos para a elaboração do mapa conceitual e para a criação do *Knowledge Graph*.

### 5.2.2 Ferramentas computacionais para a criação do mapa conceitual e do *Knowledge Graph* + *frames* semânticos

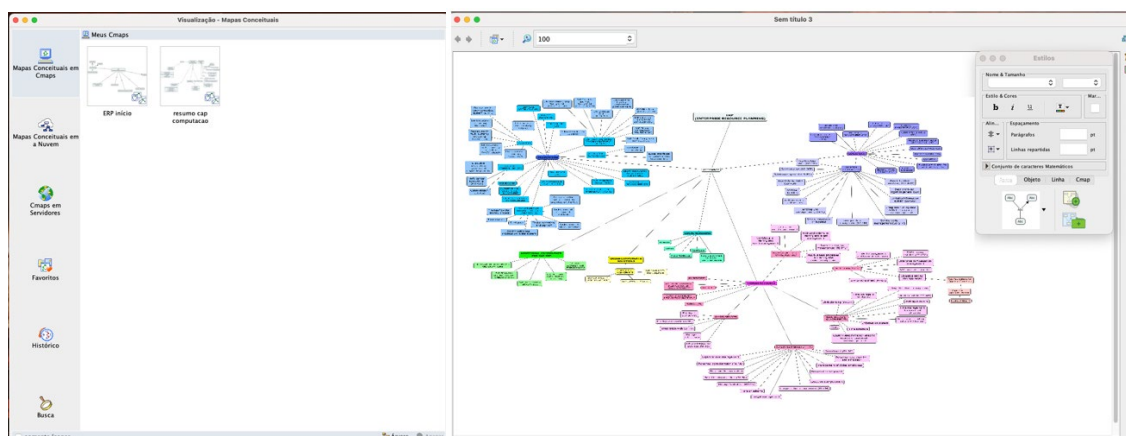
Sendo nosso objetivo elaborar um *Knowledge Graph* com *frames* semânticos para que seja utilizado tanto, de forma experimental, na base de dados da empresa como no LLM (tal qual uma das estratégias a serem adotadas nessa etapa), nesta seção, trataremos das ferramentas que utilizaremos para a criação dessa arquitetura. Inicialmente, apresentaremos a ferramenta CmapTools, a qual será empregada para criar o mapa conceitual do ERP. Em seguida, examinaremos algumas ferramentas computacionais utilizadas para elaborar ontologias e *Knowledge Graphs*, indicando quem desenvolveu e as observações sobre cada uma delas e, por fim, discorreremos sobre a ferramenta que selecionamos para a construção do nosso *Knowledge Graph* com *frames* semânticos.

Seguindo os passos metodológicos de Müller (2015), após a compilação e o processamento dos *corpora* e da aquisição do conhecimento, para a etapa de conceitualização, utilizaremos a ferramenta CmapTools<sup>51</sup> para a elaboração do mapa conceitual do ERP, como mostra a Figura 44.

---

<sup>51</sup> Ferramenta computacional disponível em: <https://cmap.ihmc.us/> Acesso em: 21 fev. 2024.

Figura 44 - Ferramenta CmapTools



Fonte: Elaborada pela autora.

Com o objetivo de sistematizar as informações obtidas nos *corpora*, utilizaremos a ferramenta *Cmap Tools*, ilustrada pela Figura 45, para a criação de mapas conceituais, com a finalidade de identificar as redes de conceitos relacionados às áreas que integram o ERP. Além disso, esse recurso computacional também será utilizado para desenvolver o mapa conceitual do ERP em sua totalidade, considerando que ele abrange todos os departamentos e, dessa forma, será possível observar as relações das áreas com o *software*. Nesse sentido, a partir da construção dos mapas conceituais específicos de cada um dos setores, também desenvolveremos um mapa geral – mapa do ERP – que integrará todos os mapas menores, assim como apresentará o domínio do programa de gerenciamento empresarial.

No que se refere às ferramentas computacionais para a construção de *Knowledge Graphs*, selecionamos alguns recursos para verificarmos qual se encaixa melhor em nosso trabalho, tendo em vista o objetivo de incluir uma camada de *frames* e informações linguísticas no KG. Para isso, elaboramos um quadro comparativo das ferramentas apresentado a seguir.

Quadro 8 - Comparativo entre as ferramentas computacionais

Ferramenta	Desenvolvedor	Observações
Stardog	Fundada por Kendall Clark, Mike Grove e Evren Sirin	<ul style="list-style-type: none"> <li>⇒ Permite acessar qualquer base de dados;</li> <li>⇒ Permite visualizar o KG online (cloud);</li> <li>⇒ Auxilia na construção do KG e resolve problemas complexos durante a elaboração;</li> <li>⇒ Unifica os dados para analisar os resultados;</li> <li>⇒ Oferece mais de um trilhão de triplas.</li> <li>⇒ Oferece apenas no formato RDF;</li> <li>⇒ Oferece o <i>software</i> de forma gratuita,</li> </ul>

		permitindo a criação de no máximo 3 KGs e de forma compartilhada.
Protégé	<i>Stanford Center for Biomedical Informatics Research at Stanford University School of Medicine.</i> Inicialmente desenvolvida por Mark Musen, em 1987.	<ul style="list-style-type: none"> <li>⇒ A Plataforma é gratuita;</li> <li>⇒ Permite construir aplicativos simples e complexos baseados em ontologia;</li> <li>⇒ Suporta as linguagens OWL e RDF.</li> </ul>
Neo4j	Empresa Neo4j, criada em 2007.	<ul style="list-style-type: none"> <li>⇒ Permite que o usuário trabalhe de forma online e offline;</li> <li>⇒ Permite acessar diversas bases de dados;</li> <li>⇒ A versão desktop é gratuita para projetos menores;</li> <li>⇒ Permite selecionar a linguagem de programação que preferir.</li> </ul>
Ontotext	A empresa iniciou em 2000, como um laboratório de pesquisa dentro do grupo Sirma. Em 2019, a empresa Sirma AI comprou a Ontotext.	<ul style="list-style-type: none"> <li>⇒ A empresa oferece uma versão gratuita e outra paga da ferramenta;</li> <li>⇒ A linguagem é RDF;</li> <li>⇒ Permite conhecer os dados com KGs já utilizados;</li> <li>⇒ Preserva toda a complexidade de modelos ontológicos;</li> <li>⇒ Desenvolve e mantém KGs de diferentes dados;</li> <li>⇒ Cria acesso de GraphQL de ontologias;</li> <li>⇒ Gera entradas SPARQL.</li> </ul>
Graphifi	Criada por Ravinder Singh e Paul Appleby, em 2019.	<ul style="list-style-type: none"> <li>⇒ Oferece a linguagem RDF;</li> <li>⇒ Suporta GraphQL;</li> <li>⇒ A plataforma oferece uma versão gratuita, com a construção de 2 produtos.</li> </ul>
Anzo	Cambridge Semantics Inc., fundada em 2007.	<ul style="list-style-type: none"> <li>⇒ A ferramenta usa a linguagem RDF;</li> <li>⇒ Permite adicionar definição, propriedades e termos às classes;</li> <li>⇒ Integra as partes das fontes de dados que apresentam vocabulários ou ontologias do KG;</li> <li>⇒ Permite harmonizar os dados de origem com o SKOS ou outras ontologias superiores;</li> <li>⇒ Oferece uma versão gratuita.</li> </ul>

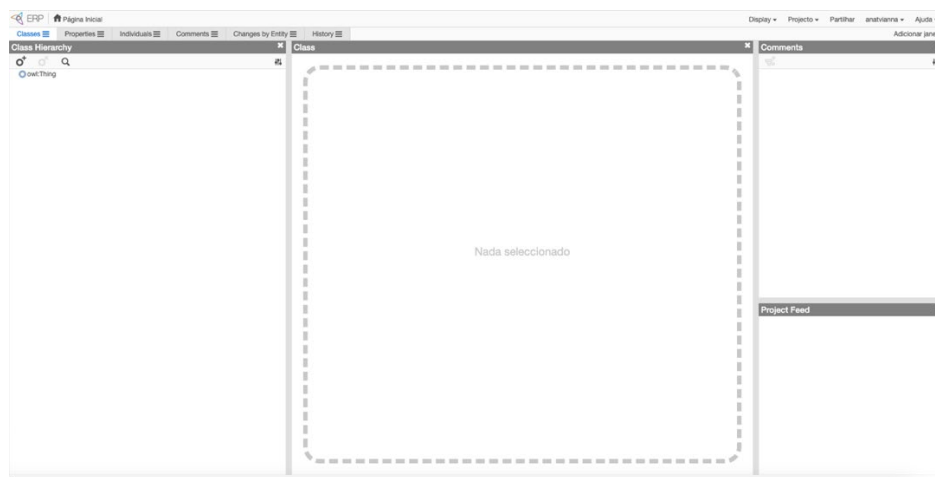
Fonte: Elaborado pela autora.

No Quadro 8, como é possível verificar, elencamos seis ferramentas para a criação de *Knowledge Graphs*: Stardog, Protégé, Neo4j, Ontotext, Graphifi e Anzo. Para cada ferramenta, buscamos os dados sobre o desenvolvedor e demais informações pertinentes sobre possibilidades de utilizá-las, como as linguagens suportadas, relações com outras ferramentas, valores de uso (gratuita ou paga), dentre outros esclarecimentos.



Atentando para os objetivos do nosso trabalho, inicialmente, selecionamos a ferramenta Protégé<sup>52</sup> para desenvolvermos nosso *Knowledge Graph*. Tal recurso computacional é desenvolvido pelo *Stanford Center for Biomedical Informatics Research*, na Escola de Medicina da Universidade de Stanford. O Protégé oferece duas versões para a criação de ontologias, a versão *online*, WebProtégé (Figura 45), e a versão *desktop* (Figura 46), sendo ambas gratuitas e abertas. As duas versões permitem que o usuário desenvolva suas arquiteturas na linguagem RDF e OWL, possibilitando a interoperabilidade entre ferramentas, uma vez que, com a facilidade de se trabalhar com as linguagens RDF e OWL, é possível baixar a arquitetura e utilizá-las em outro programa.

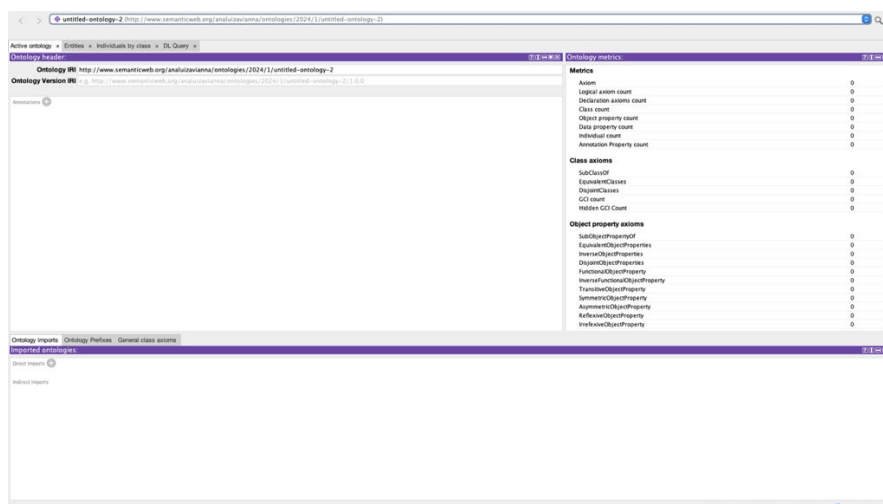
Figura 45 - Interface do WebProtégé



Fonte: Musen (2024).

---

<sup>52</sup> MUSEN, M.A. Protégé. [S.l.], 2024. Disponível em: <https://protege.stanford.edu/> Acesso em: 25 fev. 2024.

Figura 46 - Interface do Protégé versão *desktop*

Fonte: Musen (2015).

No que se refere à interface do usuário, a ferramenta apresenta as anotações para cada entidade selecionada, a descrição das classes, as hierarquias de classes, a propriedade dos dados da hierarquia de ontologias ativas, a importação de ontologias, a descrição individual da ontologia, a exibição de uma lista de instâncias, as características de propriedades de objetos, a descrição das propriedades de objetos, a *header* da ontologia – versão IRI –, as métricas da ontologia e os usos da entidade selecionada. Por se tratar de uma ferramenta desenvolvida no ambiente acadêmico, a comunidade que utiliza o *software* é muito ativa e, dessa forma, disponibiliza inúmeros fóruns para solucionar dúvidas e dificuldades na elaboração de ontologias, bem como tutoriais de como criar um grafo. Dessa forma, utilizar o Protégé é uma das melhores opções dentre as ferramentas selecionadas, sobretudo por se tratar de um recurso voltado para a academia.

Ainda que tenhamos escolhido utilizar a ferramenta *Protégé* para a elaboração do nosso *Knowledge Graph*, no ambiente de pesquisa ao qual este trabalho está vinculado, utiliza-se a ferramenta Anzo, criada pela *Cambridge Semantics Inc.*, para a representação do conhecimento. Contudo, selecionamos o Protégé, visto que ele é gratuito, permitindo utilizar todos os recursos e elaborar diversas ontologias. Para utilizar o recurso Anzo, é necessário pagar pela versão completa. Considerando isso, nossa ideia é elaborar nosso KG no *Protégé* e, como a ferramenta permite utilizar diferentes linguagens, posteriormente, poderemos inserir nosso *Knowledge Graph* na ferramenta Anzo.

Após essa revisão sobre as ontologias e *Knowledge Graphs*, a análise das ferramentas para a criação da representação do conhecimento e a seleção de qual recurso utilizaremos neste trabalho, a seguir, discorreremos sobre Engenharia de *Prompt* e os tipos de *prompt* em LLMs.

### 5.3 LLM: Estratégias de *Prompt*

Conforme discutimos no capítulo anterior, A interface linguístico-computacional: Inteligência Artificial, Processamento de Linguagem Natural e *Large Language Model*, a interação em um LLM (*Large Language Model*, grande modelo de linguagem) é feita por meio de *prompts*, isto é, o ser humano fornece instruções para que o modelo de linguagem produza uma resposta adequada à pergunta solicitada. Com o crescente desenvolvimento e o uso dos grandes modelos de linguagem, a Engenharia de *Prompt* (*Prompt Engineering*) passou a se ocupar do processo de planejar, processar e otimizar *prompts* para que o modelo de linguagem possa ser instruído e forneça uma resposta (Adelson, 2023). Atentando para o impacto que a qualidade e a eficácia dos *prompts* utilizados para o treinamento de modelos geram na performance de um LLM, é preciso refletir criticamente sobre como criar um *prompt* a fim de obter uma resposta mais precisa e útil. Nesse sentido, segundo Adelson (2023), o principal objetivo da Engenharia de *Prompt* é criar *prompts* que consigam transmitir com eficiência as entradas e saídas desejadas do modelo e, simultaneamente, minimizar a ambiguidade, o ruído e outros fatores que possam interferir na precisão e no êxito de resposta do LLM. Logo, a Engenharia de *Prompt* trabalha com a criação de *prompts*, solicitações ou instruções que orientam a saída de um modelo de linguagem. Com isso, os usuários podem controlar a saída do LLM e permitir que ele gere textos adaptados as suas necessidades (John, 2023).

Ao se trabalhar com *prompts* mais bem desenvolvidos, é possível melhorar a performance, aumentar a sua flexibilidade, aperfeiçoar a generalização e aprimorar a interpretação do modelo. Para isso, parte-se de uma fórmula de *prompt* que geralmente é composta por três elementos principais (John, 2023):

- ⇒ Tarefa: Uma declaração clara e concisa do que o *prompt* está solicitando ao modelo.
- ⇒ Instrução: A instrução deve ser seguida pelo modelo para gerar a saída.

⇒ Função: A função que o modelo deve assumir ao gerar o texto.

Ainda que tenhamos uma estrutura básica – tarefa, instrução e função – para o desenvolvimento de *prompts*, vale ressaltar que o usuário deve trabalhar suas habilidades linguísticas e expressões criativas para que as solicitações sejam ajustadas e refinadas de acordo com a tarefa a fim de se obter melhores resultados. Além disso, algumas técnicas e estratégias são utilizadas para tirar o máximo de proveito do LLM. Dentre as técnicas de *prompt* utilizadas, destacamos as seguintes estratégias: *prompt* padrão (*standard prompt*), *zero-shot prompting*, *one-shot prompting*, *few-shot prompting*, *prompt* de cadeia de pensamento (*chain-of-thought prompting*), *prompt* de árvore de pensamento (*tree of thought prompt*), *prompt* maiêutico (*maieutic prompt*), *prompt* baseado na complexidade (*complexity-based prompt*), *prompt* de conhecimento gerado (*generated knowledge prompting*), *prompt* de estímulo direcional (*directional stimulus prompting*), *prompt* com contexto (*contextual prompting*), aprendizado ativo (*active learning*), aprendizado por transferência (*transfer learning*), *data augmentation* e *prompt* passo a passo (*step-by-step prompt*).

O *prompt* padrão (*standard prompt*) é uma forma simples de guiar a saída do modelo, fornecendo uma tarefa específica para que o LLM complete a atividade (John, 2023). Esse tipo de *prompt* geralmente é combinado com outras técnicas, visto que, muitas vezes, somente uma pergunta simples não é o suficiente para que se obtenha o melhor resultado. Em seguida, temos a técnica de *zero-shot prompting*, a qual é utilizada quando não há exemplos disponíveis para a tarefa. Assim, o modelo recebe uma instrução e gera a resposta com base na sua compreensão da tarefa solicitada. Essa técnica permite que o LLM gere respostas à tarefa para as quais não foi especificamente treinado. A *one-shot prompting* é usada quando há apenas um exemplo disponível para auxiliar na tarefa. Dessa forma, o modelo recebe o exemplo e gera a resposta baseado na sua compreensão da amostra. A técnica *few-shot prompting* é empregada quando há um número limitado de exemplos disponíveis para a tarefa. Nesse sentido, o modelo recebe os exemplos e gera a resposta fundamentado na sua compreensão sobre eles (John, 2023; Mizrahi, 2023).

A técnica de *prompt* de cadeia de pensamento (*chain-of-thought prompting*) inclui uma lógica para cada exemplo, auxiliando o modelo a verbalizar as etapas intermediárias de raciocínio para resolver uma determinada tarefa. Assim, o *prompt* separa uma questão complexa em partes menores e lógicas tal qual uma linha de

pensamento (Wang *et al.*, 2023; Wei *et al.*, 2023). Na Figura 47, podemos observar a estrutura do *prompt* de cadeia de pensamento.

Figura 47 - *Prompt* de cadeia de pensamento

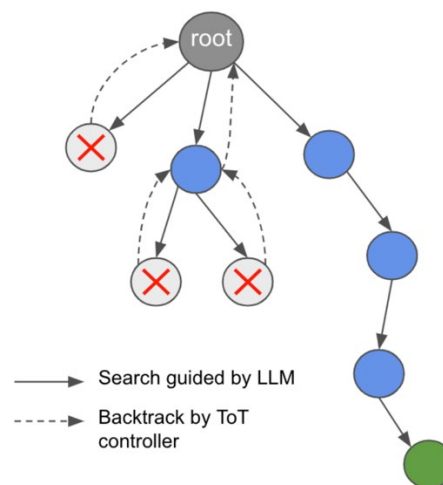
	In-context Demonstration	Inference by LLM
<b>Query</b>	Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. If she wants to read half of the remaining pages tomorrow, how many pages should she read?
<b>Standard</b>	The answer is 39.	The answer is 18. ✘
<b>CoT</b>	Originally, Leah had 32 chocolates and her sister had 42. So in total they had $32 + 42 = 74$ . After eating 35, they had $74 - 35 = 39$ pieces left in total. The answer is 39.	Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of $12 + 24 = 36$ pages. Now she has $120 - 36 = 84$ pages left. Since she wants to read half of the remaining pages, she should read $84 / 2 = 42$ pages. The answer is 42. ✔
<b>Invalid Reasoning</b>	Originally, Leah had 32 chocolates and her sister had 42. So her sister had $42 - 32 = 10$ chocolates more than Leah has. After eating 35, since $10 + 35 = 45$ , they had $45 - 6 = 39$ pieces left in total. The answer is 39.	Yesterday, Julie read 12 pages. Today, she read $12 * 2 = 24$ pages. So she read a total of $12 + 24 = 36$ pages. Now she needs to read $120 - 36 = 84$ more pages. She wants to read half of the remaining pages tomorrow, so she needs to read $84 / 2 = 42$ pages tomorrow. The answer is 42. ✔

Fonte: Wang *et al.* (2023, p. 2717).

Para a criação do *prompt* de cadeia de pensamento, é necessário apresentar dois componentes básicos: os objetos de ponte (*bridging objects*), que são os objetos principais e necessários que o modelo precisa para compreender e fazer uma previsão final bem-sucedida, como, por exemplo, números, caso seja um problema matemático, ou as entidades de sujeito e objeto, caso seja uma pergunta de controle de qualidade factual. O segundo componente se refere ao *template* de linguagem, isto é, as partes complementares dos objetos de ponte, que servem como dicas textuais e relações/predicados que orientam o LLM a compreender os objetos pontes corretos ao longo do texto. Considerando isso, o *prompt* de cadeia de pensamento irá apresentar uma descrição completa e detalhada do problema, como podemos ver na Figura 48, em que há a descrição do seguinte problema: Julie está lendo um livro de 120 páginas; ontem, ela leu 12 páginas e, hoje, Julie leu 24 páginas. Se Julie leu 36 das 120 páginas e deseja ler a metade das páginas restantes, ela deverá ler 42 páginas. Nesse exemplo, o LLM considera todo o detalhamento do *prompt* e também gera uma resposta completa, apresentando todo o raciocínio por trás da resposta. Esse tipo de técnica pode ser utilizado para solucionar problemas matemáticos, raciocínio de senso comum e manipulação simbólica, podendo – a priori – também ser aplicada a qualquer tarefa que o humano possa resolver por meio da linguagem.

O *prompt* de árvore de pensamento pode ser comparado com a forma que o ser humano soluciona seus problemas, ou seja, o homem pesquisa em um espaço combinatório por soluções parciais de um problema (Yao *et al.*, 2023). Nesse sentido, podemos representar esse tipo de pensamento como uma árvore, em que o nó raiz representa o estado inicial, o qual corresponde ao momento em que a mente humana lê a descrição do problema e inicia o processo de raciocínio. Em seguida, os outros nós representam soluções parciais e os ramos se referem aos operadores que modificam os nós. Por esse viés, o ramo a ser seguido é determinado por meio de heurísticas que auxiliam na compreensão do espaço do problema e orientam a pessoa em direção a sua solução. À luz disso, o LLM pode utilizar os nós como base para gerar a próxima etapa da pesquisa (Long, 2023). Na Figura 48, podemos visualizar como funciona a estrutura de árvore de pensamento.

Figura 48 - *Prompt* de árvore de pensamento



Fonte: Long (2023, p. 2).

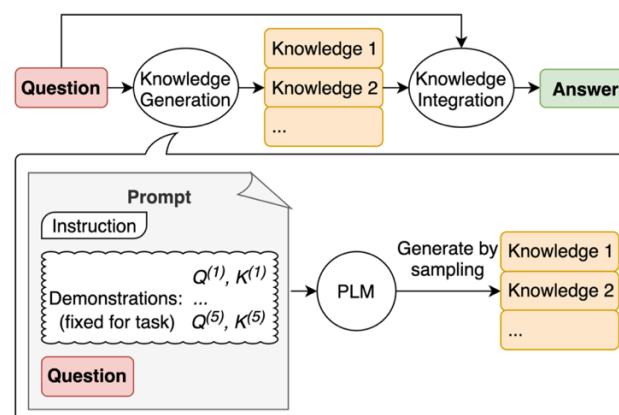
Long (2023) apresenta o nó raiz (*root*), que indica o estado inicial da árvore do pensamento e, em seguida, encontramos os demais nós, os quais se referem às soluções parciais do problema e que podem ser utilizadas pelo LLM para gerar a próxima etapa de pesquisa. Por fim, encontramos o último nó, que representa a resposta final desejada. Com relação aos ramos, podemos ter ramos que são pesquisas guiadas pelo LLM (*Search guided by LLM*) e os passos do controlador (*Backtrack by ToT controller*) da árvore de pensamento (seta pontilhada).

Similar à árvore de pensamento, o *prompt* maiêutico se fundamenta nas inferências sobre uma árvore maiêutica de explicações geradas (Jung *et al.*, 2022).

Nesse tipo de técnica, é solicitado que o modelo responda a uma determinada pergunta com uma explicação. Assim, é requisitado que o modelo explique partes da resposta, ao passo que as explicações inconsistentes são retiradas, melhorando o desempenho de raciocínio complexo de senso comum. Dessa forma, o *prompt* maiêutico faz com que o LLM gere explicações abduativas para hipóteses com raciocínio recursivo e, posteriormente, elimina as respostas contraditórias, mantendo apenas o que é consistente (Jung *et al.*, 2022).

O *prompt* baseado na complexidade é uma técnica que segue um esquema de seleção de exemplos para o raciocínio em cadeia de pensamento em várias etapas. Geralmente, essa técnica é utilizada para solucionar problemas matemáticos, uma vez que o modelo escolhe os resultados com as cadeias de pensamentos mais longas e, depois, prioriza a conclusão mais próxima do que se espera. Nesse sentido, essa técnica opta por exemplos com cadeias de raciocínio complexas e aqueles com mais etapas de raciocínio são selecionadas como *prompt* (Fei *et al.*, 2023). O *prompt* de conhecimento gerado é uma estratégia utilizada para obter informações novas e originais do modelo. De acordo com Liu *et al.* (2022), essa técnica pode gerar conhecimento útil a partir de um modelo de linguagem e, em seguida, fornecer o conhecimento como *prompt* adaptado, o qual é conectado a uma pergunta. Esse é um método simples que extrai declarações de conhecimento (expressos como declarações em linguagem natural) de modelos genéricos em uma configuração de *few-shot*. Na Figura 49, podemos ver como funciona o *prompt* de conhecimento gerado.

Figura 49 - Processo de *prompt* de conhecimento gerado



Fonte: Liu *et al.* (2022, p.1).

Como podemos notar na Figura 49, tal técnica consiste em dar uma instrução com algumas demonstrações fixadas para cada tarefa e com um espaço para novas

perguntas. As demonstrações são escritas por humanos, e cada uma consiste em uma pergunta no estilo da tarefa e uma declaração de conhecimento adequada para responder tal pergunta (Liu *et al.*, 2022).

A técnica de *prompt* de estímulo direcional (*Directional stimulus prompting*) fornece dicas e pistas para que a consulta de entrada guie o LLM em direção ao resultado desejado (Li *et al.*, 2023). Dessa forma, a técnica passa a incorporar palavras-chave no *prompt* como um estímulo direcional para indicar os principais aspectos que o resultado desejado deve cobrir. A título de exemplo, para solicitar que o LLM elabore um poema sobre amor, na instrução da tarefa – seguindo a técnica de estímulo direcional –, podemos incluir as palavras-chave “amor”, “paixão” e “eterno” como dicas para obtermos uma resposta mais precisa. Conforme Li *et al.* (2023) afirmam, ao inserir essa orientação no *prompt* com o estímulo direcional, o modelo pode gerar resultados mais próximos do esperado na referência.

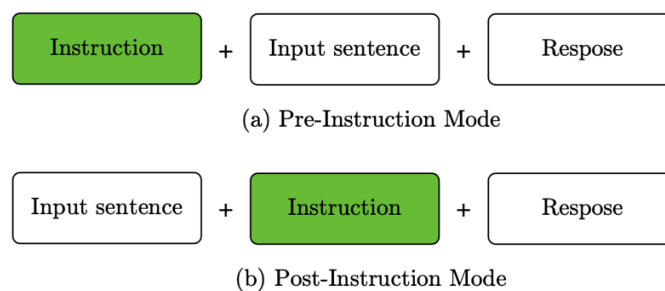
Além das técnicas já descritas, Mizrahi (2023) destaca que também podemos encontrar *prompts* com adição de contexto (*contextual prompting*), em que o usuário inclui informações contextuais para orientar a resposta do modelo. Essa técnica auxilia o LLM a compreender qual é a intenção do usuário, o formato desejado da saída ou quaisquer outras restrições que devem ser levadas em conta ao gerar uma resposta. Outra possibilidade é a técnica de aprendizado ativo (*active learning*), no qual o LLM recebe uma pequena quantidade de dados rotulados e, depois, identifica os exemplos mais informativos a serem anotados. Ainda temos o aprendizado por transferência (*transfer learning*), em que o LLM é treinado em uma tarefa semelhante à atividade que lhe será solicitado a executar. Além dessas, a técnica de *data augmentation* consiste em treinar o LLM com um conjunto de dados que foi criado artificialmente pela adição de ruído ou variações aos dados existentes. Ao aplicar essa técnica, o modelo pode melhorar seu desempenho em tarefas com poucos dados rotulados disponíveis. Por fim, Mizrahi (2023) apresenta a técnica de *prompt* passo a passo (*step-by-step prompt*), a qual exige que o modelo forneça uma sequência de etapas ou um procedimento para realizar uma determinada tarefa, como uma receita, um tutorial ou um algoritmo, por exemplo. Normalmente, essa técnica é aplicada à resolução de problemas matemáticos.

Ainda que tenhamos diferentes técnicas de *prompt* para realizar experimentos e obter os resultados mais eficientes e precisos, Adelson (2023) salienta que é preciso atentar para alguns pontos quando for desenvolver um *prompt*. Segundo o autor, é necessário (i) ter um objetivo bem definido, (ii) usar contexto e palavras-chave



relevantes, (iii) ser específico e conciso, (iv) experimentar diferentes tamanhos de *prompt* e diferentes complexidades na instrução, (v) avaliar os *prompts* gerados de forma crítica e (vi) repetir e redefinir o processo de criação de *prompt*. Esses passos auxiliam o usuário a obter respostas mais apuradas, precisas e úteis para a sua tarefa. Além de considerarmos esses pontos para a criação de *prompt*, também é preciso refletir sobre a disposição da informação. Como vimos no capítulo anterior, o estudo de Liu *et al.* (2023) mostrou que a sequência de instrução no *prompt* deve ser bem elaborada para que se obtenha uma resposta mais acurada. Os autores apontam que o tamanho da instrução e a posição das informações são fundamentais na hora da criação. Nesse sentido, Liu *et al.* (2023) mostram que instruções longas e desconexas não orientam o LLM a produzir uma resposta adequada a determinada tarefa. Logo, eles sugerem a seguinte estrutura de instrução, conforme Figura 50.

Figura 50 - Modelo de instrução de *prompt*



Fonte: Liu *et al.* (2023).

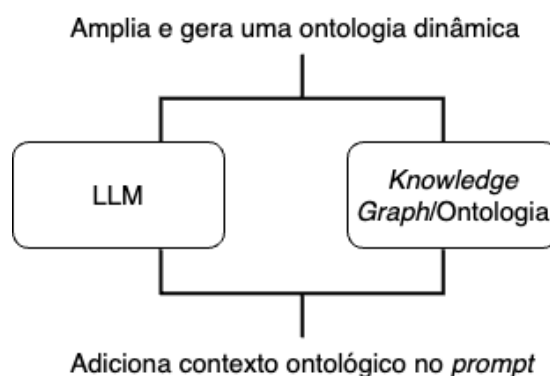
Geralmente, os usuários criam *prompts* considerando a primeira forma de instrução *Pre-Instruction Mode* (Modo pré-instrução): a) instrução + sentença de entrada + resposta correspondente. Essa estrutura padrão de instrução pode levar o LLM a esquecer alguns aspectos importantes da solicitação, visto que esse formato é mais longo, e o LLM pode não captar todas as informações. A abordagem (*Post-Instruction Mode* – Modo Pós-Instrução) sugerida pelos autores é composta por b) sentença de entrada + instrução + resposta correspondente. Nessa estrutura, realocamos a instrução para depois da sentença de entrada. Essa mudança estimula o LLM a fornecer respostas mais completas e precisas (Liu *et al.*, 2023).

Outra questão já discutida neste trabalho diz respeito à utilização da técnica de *prompt* com a inclusão de representação do conhecimento. Dentre alguns dos motivos que levam a buscar subsídios na representação do conhecimento como técnica para aprimorar os resultados de um LLM estão as alucinações. Por apresentarem limitações

quanto às habilidades de raciocinar acerca de relações complexas entre entidades e conceitos, as ontologias e os *Knowledge Graphs* podem auxiliar os LLMs ao fornecerem conhecimento externo que será utilizado para que o modelo raciocine e gere textos mais acurados. Nessa aproximação entre KG, ontologias e LLMs, é possível aprimorar a competência de realizar perguntas mais complexas que exigem um conhecimento prévio do LLM. Nesse sentido, o modelo de linguagem pode usar o KG para identificar entidades e relações mais relevantes e, assim, gerar respostas mais precisas e úteis. Além disso, pode-se utilizar o LLM para enriquecer um *Knowledge Graph* ao extrair e conectar entidades de um texto não estruturado (Moiseev *et al.*, 2022; Goyal; Bhattacharyya, 2023).

O uso de técnicas de representação do conhecimento para melhorar o uso dos LLMs pode ser visto sob duas perspectivas. De um lado, o modelo de linguagem pode contribuir para desenvolver melhorias em ontologias e *Knowledge Graph*, e, por outro lado, a representação do conhecimento pode ser utilizada no LLM para melhorar sua habilidade de raciocinar e fornecer respostas mais precisas. Na Figura 51, podemos observar essa relação entre LLM e *Knowledge Graphs*.

Figura 51 - Relação entre LLM e *Knowledge Graph*



Fonte: Elaborada pela autora.

Partindo do ponto de vista do LLM, o modelo pode contribuir para a extração de informações de textos e auxiliar no mapeamento da ontologia, criar listas iniciais de conceitos para serem utilizados como ponto de partida de uma ontologia ou instâncias de um KG, gerar entradas SPARQL de linguagem natural, resumir uma ontologia ou um *Knowledge Graph*, ajudar no alinhamento de ontologias e KGs, dentre outras possibilidades. Sob outro enfoque, as ontologias e *Knowledge Graphs* podem validar as respostas do LLM, fornecer dados de treinamento ou entrada de *prompts* para o modelo,

auxiliar o raciocínio lógico e inferencial, dentre outras contribuições. Nesse sentido, o uso de representação do conhecimento como técnica de *prompt* em LLM é fundamental para que se ajuste o modelo e o treine, a fim de obter respostas mais próximas das desejadas.

Considerando os pontos discutidos acerca das técnicas de *prompt* e como utilizá-las, na próxima seção, apresentaremos as estratégias aplicadas para a tarefa de harmonização terminológica com LLM.

### 5.3.1 Terminologia e LLMs: Instruções e estratégias de *prompt* para a tarefa de harmonização terminológica

Atentando para o nosso objetivo de explorar técnicas de *prompt* para trabalhar a harmonização terminológica de forma automática, utilizando o ChatGPT, nesta seção, apresentaremos a metodologia aplicada para a solicitar a tarefa de harmonizar termos e conceitos. Para isso, empregamos diferentes estratégias, a fim de compreender o funcionamento do LLM, bem como examinar os resultados produzidos. Nesse sentido, selecionamos diferentes técnicas, mas, mais do que isso, instruímos o modelo em inglês e em português para investigarmos se as respostas seriam iguais nas duas línguas. Como nosso *corpus* é em inglês e utilizaremos a base de dados da empresa para buscarmos as traduções para o português, torna-se importante, e ao mesmo tempo interessante, trabalharmos com uma etapa de tradução e análise contrastiva no LLM. Esse procedimento metodológico poderá apresentar evidências quanto à construção dos *frames* e a relevância de se considerar essa camada semântica tanto para a análise da terminologia e da tradução como para a implementação computacional, visto que não conceptualizamos e estruturamos as cenas e eventos da mesma forma em todas as línguas.

Antes de apresentarmos as perguntas e analisarmos como foram empregadas as técnicas e estratégias de *prompt*, é fundamental tratarmos da definição dos parâmetros do LLM. O primeiro passo antes de solicitar alguma tarefa é ajustar os parâmetros do modelo de linguagem. Conforme afirmam Kublik e Saboo (2022), os LLMs utilizam abordagens probabilísticas em vez de lógica convencional. Logo, eles podem gerar uma vasta gama de respostas para uma mesma entrada, dependendo, sobretudo, de como os parâmetros do modelo são definidos. A depender dessa definição, um LLM busca

encontrar a melhor correspondência probabilística dentro do conjunto de dados em que ele foi treinado.

Entre os parâmetros a serem definidos, encontramos a temperatura (*temperature*), fator que controla a aleatoriedade da resposta, sendo representada como um intervalo entre 0 e 1. A temperatura está relacionada à criatividade da resposta dada pelo LLM, isto é, se o valor da temperatura for baixo, considerando o intervalo entre 0 e 1, o modelo fornecerá uma resposta com o primeiro resultado que encontrar, gerando um texto mais correto e com pouca variação. Caso a temperatura seja definida como alta, o modelo avaliará as possíveis respostas que se encaixam melhor no contexto, antes de prever o resultado. Ainda que, com essa abordagem, o LLM possa gerar resultados mais diversificados, há grandes chances de fornecer um texto com muitos erros gramaticais e alucinações (Kublik; Saboo, 2022). Para o nosso trabalho, utilizamos dois modelos de GPT – GPT-4 e GPT-4 32k e definimos os seguintes parâmetros, conforme mostra a Figura 52.

Figura 52 - Definição dos parâmetros GPT

The figure displays three screenshots of the GPT parameter configuration interface, each for a different model. Each interface includes a dropdown menu for 'Choose an Example', a dropdown for the model name, and buttons for 'Show/Hide Parameters' and 'Reset'. The parameters are adjusted using sliders.

Model	Max Tokens	Temperature	Frequency Penalty	Presence Penalty
Azure OpenAI – gpt-4-32k	8900	0	0	0
Azure OpenAI – gpt-4	4000	0	0	0
Azure OpenAI – gpt-4	4000	1	0	0

Fonte: OpenAI (2024).

Como podemos observar na Figura 52, definimos diferentes parâmetros, a fim de comparar as respostas obtidas. Para o modelo GPT-4 32k, opta-se pela temperatura em zero. Para o modelo GPT-4 32k, definimos duas temperaturas, zero e um. Escolhemos trabalhar com as duas temperaturas no modelo GPT-4 32k para observarmos se haverá variações nas respostas, uma vez que a temperatura em zero apresenta um texto mais correto e com poucas variações, e o modelo com a temperatura em 1 apresenta um texto mais criativo e com a possibilidade de alucinar. O GPT-4 é a última criação da OpenAI. Ele foi lançado em março de 2023 e é um modelo multimodal. O GPT-4 32k aceita uma quantidade maior de *tokens*, podendo gerar textos mais longos. Tais modelos não estão disponíveis gratuitamente na plataforma da OpenAI. No que se refere à quantidade de *tokens*, o modelo GPT-4 32k oferece a quantidade máxima de 32.768 *tokens*, enquanto o modelo GPT-4 oferece o máximo de 4 mil *tokens*. Em um *Large Language Model*, os *tokens* pode ser identificados como palavras inteiras, partes de palavras ou até caracteres. De acordo com Ozdemir (2023), o *token* condiz com a menor unidade de significado criado pela divisão de uma frase ou trecho de texto em unidades menores, servindo como entrada para um LLM. Para este trabalho, selecionamos como quantidade máxima 8.900 *tokens* no GPT-4 32k. Isso ocorreu pela disponibilidade de *tokens* no momento para desempenhar tal tarefa, não sendo possível selecionar a quantidade máxima do modelo, uma vez que nosso acesso aos modelos é limitado.

Para a elaboração das instruções a serem inseridas no *prompt* como estratégia para a tarefa de harmonização automática, julgamos ser necessário iniciar com uma pergunta mais abrangente, averiguando os conhecimentos prévios do modelo acerca da harmonização de termos e conceitos para, finalmente, desenvolvermos *prompts* mais complexos. Nesse sentido, inicialmente, solicitamos ao ChatGPT que respondesse o que é a harmonização terminológica. Em um segundo momento, pedimos que ele realizasse a harmonização dos termos, apresentando o termo principal e as variantes, sem a adição de contexto ou outras informações que pudessem interferir na tarefa. A terceira instrução foi a solicitação da harmonização do conjunto de termos, considerando as definições da terminologia para a tarefa. Para essa etapa, inserimos no *prompt* as definições dos termos coletados na base de dados da empresa<sup>53</sup>, sem incluir exemplos nas definições. A quarta solicitação foi para que o LLM fornecesse a harmonização do

---

<sup>53</sup> É importante salientar que nossos *corpora* foram elaborados em língua inglesa. À vista disso, para as instruções em língua portuguesa, buscamos as traduções dos termos na base de dados da empresa. As definições foram traduzidas despretensiosamente por nós, em função da base de dados da empresa não apresentar a tradução das definições.

conjunto de termos, considerando as definições coletadas da base de dados da empresa, juntamente com alguns exemplos. Posteriormente, instruímos o *prompt* à harmonização do conjunto de termos, considerando a verbalização do nosso *Knowledge Graph* com *frames* semânticos. Por fim, o sexto *prompt* solicitado se referiu à harmonização dos termos conforme as definições da base de dados, os exemplos e a verbalização. Podemos observar as seis instruções no Quadro 9.

Quadro 9 - Instruções solicitadas ao LLM em língua portuguesa

Instrução 1: O que é harmonização terminológica?
Instrução 2: Considerando a harmonização terminológica, como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal. Qual seria o termo principal e quais seriam as variantes?
Instrução 3: Considere as seguintes definições: Uma pessoa que contribui ou contribuiu para a criação de bens e serviços na empresa com base em um contrato de trabalho, um acordo sem aviso prévio ou um contrato de serviços. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços para uma empresa. O funcionário inclui funcionários internos e externos (executores de serviços). Diferentemente dos externos, o funcionário interno está vinculado a instruções e é obrigado a aderir às políticas e aos regulamentos da empresa. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços na empresa com base em um contrato de trabalho para serviços. Um funcionário que não tem um contrato explícito de emprego permanente ou de longo prazo. Um indivíduo que realiza trabalhos para uma organização em caráter não permanente. Um usuário corporativo que é um funcionário terceirizado contratado como funcionário externo. Um usuário corporativo que é um funcionário terceirizado que participa de um cenário de colaboração. Um elemento de recurso que pode operar um dispositivo ou trabalhar sem um dispositivo. Funcionário ou trabalhador externo. Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando as definições acima? Qual seria o termo principal e quais seriam as variantes?
Instrução 4: Considere as seguintes definições: Contrato de trabalho, um acordo sem aviso prévio ou um contrato de serviços. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços para uma empresa. O funcionário inclui funcionários internos e externos (executores de serviços). Diferentemente dos externos, o funcionário interno está vinculado a instruções e é obrigado a aderir às políticas e aos regulamentos da empresa. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços na empresa com base em um contrato de trabalho para serviços. O funcionário constitui o principal foco de interesse da maioria dos componentes de aplicação de recursos humanos, como administração de pessoal, gerenciamento de tempo, folha de pagamento e assim por diante. Todos os processos de negócios nesses componentes de aplicação estão relacionados ao funcionário. Um funcionário que não tem um contrato explícito de emprego permanente ou de longo

prazo.

Um indivíduo que trabalha para uma organização de forma não permanente. Os trabalhadores temporários são geralmente conhecidos como profissionais independentes, trabalhadores com contrato temporário, empreiteiros ou consultores.

Um usuário comercial que é um funcionário terceirizado contratado como funcionário externo. Por exemplo, consultores ou auditores externos. Portanto, eles compartilham o endereço de trabalho do empregador contratado por um determinado período de tempo.

Um usuário corporativo que é um funcionário terceirizado que participa de um cenário de colaboração, por exemplo, inventário gerenciado pelo fornecedor.

Um usuário de colaboração precisa de acesso a um sistema, por exemplo, para manter estoques.

Um elemento de recurso que pode operar um dispositivo ou trabalhar sem um dispositivo. Exemplo: motorista de empilhadeira ou selecionador. um operador no chão de fábrica que monta placas, unidades ou outros materiais que você fabrica.

Funcionário ou trabalhador externo.

Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando as definições acima? Qual seria o termo principal e quais seriam as variantes?

Instrução 5:

Considere a seguinte verbalização:

ERP é um software de gerenciamento de empresas.

ERP tem cross components.

Cross components tem o termo Employee (empregado).

ERP tem componentes.

Human Resources é um componente do ERP.

Human Resources tem recursos.

Payroll, Personnel Management e HR Master Data são recursos de Human Resources.

Payroll tem o termo Worker (operário).

HR Master Data tem as instâncias HCM\_Employee\_Central e Business\_User.

A instância HCM\_Employee\_Central tem os termos Contingent Worker (trabalhador temporário) e Employee (colaborador).

A instância Business User tem os termos Collaboration User (usuário de colaboração), External Resource (recurso externo) e Worker (trabalhador).

Personnel Management tem termos.

Employee (empregado) é um termo de Personnel Management.

Personnel Management tem funcionalidades.

Budget Management e Personnel Administration são funcionalidades de Personnel Management.

Budget Management tem o termo Employee (empregado).

Personnel Administration tem os termos Contingent Worker (trabalhador temporário) e External Resource (recurso externo).

ERP inclui Other Sectors.

Other Sectors tem Industries.

Industries tem Public Sector.

Public Sector tem recursos.

Public Sector Management é um recurso de Public Sector.

Investigative Case Management é uma funcionalidade de Public Sector Management.

Investigative Case Management tem o termo Employee (empregado).

Supply Chain é um componente do ERP.

Supply Chain tem recursos.

Logistic Cross Topics é um recurso de Supply Chain.

Logistic Cross Topics tem funcionalidades.

Logistic Execution é uma funcionalidade de Logistic Cross Topics.

Logistic Execution tem app para gerenciamento de Task and Resource Management

Task and Resource Management tem o termo Worker (funcionário).  
 Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando a verbalização acima. Qual seria o termo principal e quais seriam as variantes?

Instrução 6:

Considere as seguintes definições:

Empregado é uma pessoa que contribui ou contribuiu para a criação de bens e serviços na empresa com base em um contrato de trabalho, um acordo sem aviso prévio ou um contrato de serviços.

O funcionário inclui funcionários internos e externos (executores de serviços). Diferentemente dos externos, o funcionário interno está vinculado a instruções e é obrigado a aderir às políticas e aos regulamentos da empresa. O funcionário constitui o principal foco de interesse da maioria dos componentes de aplicação de recursos humanos, como administração de pessoal, gerenciamento de tempo, folha de pagamento e assim por diante.

Os trabalhadores temporários são geralmente conhecidos como profissionais independentes, trabalhadores com contrato temporário, empreiteiros ou consultores.

Um recurso externo é um funcionário terceirizado contratado como funcionário externo. Por exemplo, consultores ou auditores externos. Portanto, eles compartilham o endereço de trabalho do empregador contratado por um determinado período de tempo.

Um usuário colaboração que é um funcionário terceirizado que participa de um cenário de colaboração, por exemplo, inventário gerenciado pelo fornecedor.

Um operário pode operar um dispositivo ou trabalhar sem um dispositivo. Exemplo: motorista de empilhadeira ou selecionador. Um operador no chão de fábrica que monta placas, unidades ou outros materiais que você fabrica.

Considere a seguinte verbalização:

ERP é um software de gerenciamento de empresas.

ERP tem cross components.

Cross components tem o termo Employee (empregado).

ERP tem componentes.

Human Resources é um componente do ERP.

Human Resources tem recursos.

Payroll, Personnel Management e HR Master Data são recursos de Human Resources.

Payroll tem o termo Worker (operário).

HR Master Data tem as instâncias HCM\_Employee\_Central e Business\_User.

A instância HCM\_Employee\_Central tem os termos Contingent Worker (trabalhador temporário) e Employee (colaborador).

A instância Business User tem os termos Collaboration User (usuário de colaboração), External Resource (recurso externo) e Worker (trabalhador).

Personnel Management tem termos.

Employee (empregado) é um termo de Personnel Management.

Personnel Management tem funcionalidades.

Budget Management e Personnel Administration são funcionalidades de Personnel Management.

Budget Management tem o termo Employee (empregado).

Personnel Administration tem os termos Contingent Worker (trabalhador temporário) e External Resource (recurso externo).

ERP inclui Other Sectors.

Other Sectors tem Industries.

Industries tem Public Sector.

Public Sector tem recursos.

Public Sector Management é um recurso de Public Sector.

Investigative Case Management é uma funcionalidade de Public Sector Management.

Investigative Case Management tem o termo Employee (empregado).

Supply Chain é um componente do ERP.



Supply Chain tem recursos.  
 Logistic Cross Topics é um recurso de Supply Chain.  
 Logistic Cross Topics tem funcionalidades.  
 Logistic Execution é uma funcionalidade de Logistic Cross Topics.  
 Logistic Execution tem app para gerenciamento de Task and Resource Management  
 Task and Resource Management tem o termo Worker (funcionário).  
 Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando as definições e a verbalização acima. Qual seria o termo principal e quais seriam as variantes?

Fonte: Elaborado pela autora.

Acerca das instruções de *prompt*, como já apresentamos, estruturamos seis perguntas para desenvolvermos a tarefa de harmonização terminológica no LLM. A fim de demonstrarmos o funcionamento do modelo GPT-4 32k, elaboramos quatro *prompts* para verificarmos o conhecimento prévio do modelo acerca da tarefa de harmonização. Para isso, iniciamos nossa solicitação perguntando o que é a harmonização terminológica, conforme ilustrado pela Figura 53.

Figura 53 - *Prompt* GPT-4 32k

The image shows a screenshot of the OpenAI GPT-4 32k interface. At the top, there is a dropdown menu for 'Choose an Example' and another for 'Azure OpenAI - gpt-4-32k'. To the right are buttons for 'Show/Hide Parameters' and 'Reset'. The main input area contains the text 'O que é harmonização terminológica?'. Below the input area, there is a 'Run (CTRL+ENTER)' button and a checkbox labeled 'Enable Prompt Logging (Deactivate for sending confidential data)' which is checked.

Fonte: OpenAI (2024).

No que se refere à instrução 1, pedimos para que o LLM definisse a tarefa de harmonização. Essa solicitação foi feita para verificarmos se o modelo entende o que é harmonização terminológica, forneça exemplos de harmonização e apresente qual abordagem é utilizada para harmonizar os termos. Abaixo, podemos observar as respostas obtidas no LLM sobre harmonização terminológica (Quadro 10).

Quadro 10 - Instrução 1 e Resposta do modelo GPT-4 32k

O que é harmonização terminológica?

Harmonização terminológica é o processo de padronização e unificação de termos dentro de um determinado campo de estudo ou indústria. Isso é feito para evitar confusão e mal-entendidos que podem surgir devido ao uso de diferentes termos para se referir ao mesmo conceito.

Por exemplo, na área médica, é crucial que todos os profissionais usem a mesma terminologia ao se referir a uma determinada doença ou procedimento. Isso garante que todos entendam exatamente o que está sendo discutido, independentemente de onde estejam no mundo.

A harmonização terminológica é especialmente importante em contextos internacionais e multilíngues, onde a tradução de termos pode levar a ainda mais confusão. Organizações internacionais muitas vezes têm comitês dedicados à harmonização terminológica para garantir a consistência em todos os idiomas.

Além disso, a harmonização terminológica também é importante para a inteligência artificial e o processamento de linguagem natural, pois ajuda a treinar modelos de linguagem para entender e usar a terminologia corretamente.

Fonte: OpenAI (2024).

Com relação à definição de harmonização terminológica, o modelo afirma que essa tarefa se refere ao processo de padronizar e unificar termos de um determinado domínio, com a finalidade de evitar inconsistências e confusões que possam afetar a comunicação. Para exemplificar a tarefa, o LLM apresenta a harmonização de termos na área médica, em organizações internacionais e na área computacional, como na Inteligência Artificial e no Processamento de Linguagem Natural. Após esse questionamento inicial, perguntamos ao LLM, novamente, o que é harmonização e adicionamos a mais uma questão: Qual é a metodologia utilizada para harmonizar termos? (Quadro 11).

Quadro 11 - Instrução 2 e resposta do modelo GPT-4 32k

O que é harmonização terminológica? Qual é a metodologia utilizada para harmonizar termos?

A harmonização terminológica é um processo que busca estabelecer consistência e uniformidade no uso de termos e conceitos dentro de um determinado campo de conhecimento, indústria ou organização. Isso é especialmente importante em contextos multilíngues e internacionais, onde a falta de consistência pode levar a mal-entendidos e erros de comunicação.

A metodologia para harmonizar termos pode variar dependendo do contexto, mas geralmente envolve as seguintes etapas:

1. **Identificação de termos**: O primeiro passo é identificar os termos que precisam ser harmonizados. Isso pode envolver a revisão de documentos, manuais, diretrizes, etc., para identificar termos e conceitos chave.
2. **Análise de termos**: Uma vez identificados os termos, eles são analisados para entender seu significado, uso e contexto. Isso pode envolver a consulta a especialistas no campo, a revisão de literatura relevante, etc.

3. **\*\*Seleção de termos preferenciais\*\***: Com base na análise, são selecionados termos preferenciais para cada conceito. Esses termos são aqueles que serão usados consistentemente em toda a organização ou campo de conhecimento.
4. **\*\*Criação de um glossário ou banco de dados terminológico\*\***: Os termos preferenciais e suas definições são então compilados em um glossário ou banco de dados terminológico. Este recurso serve como uma referência para todos os que usam os termos.
5. **\*\*Implementação e monitoramento\*\***: O último passo é implementar o uso dos termos preferenciais e monitorar sua adoção. Isso pode envolver a formação de pessoal, a revisão de documentos e comunicações para garantir a consistência, etc.

A harmonização terminológica é um processo contínuo que requer revisão e atualização regulares para garantir que os termos permaneçam relevantes e atualizados.

Fonte: OpenAI (2024).

No Quadro 11, podemos vislumbrar os passos que o LLM fornece para realizar a tarefa de harmonização. Se observarmos com atenção, podemos perceber que os quatro primeiros passos que o LLM apresenta são as mesmas etapas de harmonização propostas pela ISO<sup>54</sup>. Já o quinto passo é um adicional que o modelo oferece.

Ainda que nossa primeira instrução esteja focada em solicitar que o LLM defina o que é harmonização terminológica, é interessante nos aprofundarmos nesse ponto. Pensando nisso, investigamos um pouco mais acerca dos conhecimentos sobre a tarefa e, para isso, fizemos a seguinte pergunta ao LLM (Quadro 12).

Quadro 12 - Instrução 3 e resposta do modelo GPT-4 32k

Quem propôs a harmonização terminológica?

A harmonização terminológica é um conceito que não é atribuído a uma única pessoa ou entidade. É um processo contínuo que envolve muitos profissionais e organizações ao longo do tempo. A harmonização terminológica é um esforço para garantir que termos e conceitos sejam usados de maneira consistente e precisa em diferentes contextos, especialmente em campos como a tradução, a localização e a comunicação intercultural. Organizações internacionais como a UNESCO e a ISO têm comitês dedicados à harmonização terminológica. Além disso, muitos linguistas e tradutores contribuem para esse processo em seu trabalho diário.

Fonte: OpenAI (2024).

Como podemos observar, o modelo indica que a tarefa de harmonização foi desenvolvida por organizações internacionais, como a ISO (o que comprova os passos metodológicos apresentados acima pelo modelo), assim como a UNESCO, as quais têm comitês específicos para refletir e propor tal atividade terminológica. Por fim,

<sup>54</sup> Na seção 3.3.1 Harmonização de conceitos e termos: metodologia e aplicação apresentamos o fluxograma proposto pela ISO para a harmonização de termos e conceitos.

solicitamos ao LLM que respondesse à pergunta relativa aos parâmetros para harmonizar termos:

### Quadro 13 - Instrução 4 e resposta do modelo GPT-4 32k

<p>Quais parâmetros devemos seguir para realizar a harmonização terminológica?</p> <p>A harmonização terminológica é um processo que visa garantir a consistência e a precisão do uso de termos em um determinado contexto, como em um projeto de tradução, um documento técnico ou um sistema de informação. Aqui estão alguns parâmetros que você pode seguir para realizar a harmonização terminológica:</p> <ol style="list-style-type: none"> <li>1. <b>Definição clara de termos</b>: Cada termo deve ter uma definição clara e precisa. Isso ajuda a evitar confusões e mal-entendidos.</li> <li>2. <b>Consistência</b>: O mesmo termo deve ser usado para se referir ao mesmo conceito em todo o documento ou sistema. Da mesma forma, diferentes termos não devem ser usados para se referir ao mesmo conceito.</li> <li>3. <b>Precisão</b>: Os termos devem ser precisos e refletir corretamente o conceito que representam. Isso é especialmente importante em campos técnicos e científicos, onde a precisão é crucial.</li> <li>4. <b>Contexto</b>: O uso de termos deve ser apropriado para o contexto. Por exemplo, um termo que é apropriado em um contexto técnico pode não ser apropriado em um contexto não técnico.</li> <li>5. <b>Conformidade com padrões existentes</b>: Se houver padrões existentes para a terminologia em um determinado campo, esses padrões devem ser seguidos.</li> <li>6. <b>Revisão e atualização regulares</b>: A terminologia pode mudar com o tempo, por isso é importante revisar e atualizar regularmente a terminologia para garantir que ela permaneça precisa e relevante.</li> <li>7. <b>Participação dos stakeholders</b>: Os stakeholders, como especialistas no assunto, usuários finais e tradutores, devem estar envolvidos no processo de harmonização terminológica para garantir que a terminologia atenda às suas necessidades.</li> <li>8. <b>Documentação</b>: Todo o processo de harmonização terminológica deve ser documentado. Isso inclui a definição de termos, as decisões tomadas durante o processo e quaisquer mudanças feitas na terminologia.</li> <li>9. <b>Treinamento e educação</b>: Aqueles que usarão a terminologia devem ser treinados e educados sobre a terminologia e seu uso correto.</li> <li>10. <b>Uso de ferramentas de gestão terminológica</b>: Ferramentas de gestão terminológica podem ajudar a gerir e manter a terminologia, garantindo a consistência e a precisão.</li> </ol>
---

Fonte: OpenAI (2024).

Com relação à última instrução, podemos perceber que o modelo traz mais informações sobre a harmonização, mas não extrapola as bases da ISO, como, por exemplo, introduzir uma etapa para tratar de fatores extralinguísticos na tarefa, conforme a nossa proposta semântico-terminológica. Isso indica que nossa proposta metodológica é inovadora, visto que as atuais abordagens ainda não consideram os aspectos contextuais, cognitivos, experienciais e situacionais na sua execução.

Após esse estudo inicial acerca da definição de harmonização fornecida pelo LLM, na segunda instrução, pedimos que o modelo harmonizasse um conjunto de termos composto por: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal. Para essa etapa, o modelo deveria selecionar o termo principal e as variantes. Na terceira instrução, adicionamos as definições<sup>55</sup> dos termos que compõem o conjunto de dados e pedimos que o modelo harmonizasse a terminologia considerando as definições apresentadas. Por fim, na quarta instrução, além das definições, também incluímos alguns exemplos de profissões (sublinhado) que estão relacionados à terminologia. A partir disso, o modelo deveria apresentar o termo principal e as variantes.

Embora nosso trabalho busque harmonizar os termos em língua portuguesa, realizamos alguns experimentos em língua inglesa. Essa escolha se dá por dois motivos. O primeiro se refere à necessidade de buscarmos as traduções dos termos em português, uma vez que os nossos *corpora* são compostos por textos em língua inglesa. Nesse sentido, torna-se interessante apresentar as duas perspectivas de harmonização. O segundo motivo está relacionado à estrutura de *frames*. Pelo viés da Semântica de *Frames*, sabe-se que os falantes de uma língua não ativam necessariamente o mesmo *frame* que os falantes de uma outra língua, visto que a estrutura de *frames* é concebida a partir das experiências mundanas, sociais, culturais e históricas (Fillmore, 1982). Logo, a arquitetura de *frames* de um determinado par linguístico não será elaborada da mesma forma e com as mesmas relações (Chishman; Bertoldi, 2013; Boas, 2013). Desse modo, a harmonização pode ser diferente em língua portuguesa e em língua inglesa. Considerando isso, elaboramos as mesmas instruções em inglês<sup>56</sup>, conforme apresenta o Quadro 14.

Quadro 14 - Instruções solicitadas ao LLM em língua inglesa

Instruction 1: What is terminological harmonization?
Instruction 2: Considering the terminological harmonization, how could I harmonize the following terms: workforce person, worker, collaboration user, external resource, contingent worker, and employee. What would be the main term and what would be the variants?
Instruction 3: Consider the following definitions: A person who contributes or has contributed to the creation of goods and services in the company based on a work contract, an at-will arrangement, or a contract for services. A person who contributes or has contributed to the creation of goods or services for a company.

<sup>55</sup> As definições para essa tarefa foram coletadas da base de dados da empresa. No que tange às definições em língua portuguesa, traduzimos de forma despretensiosa para que a tarefa fosse realizada em português.

<sup>56</sup> As definições coletadas para a tarefa de harmonização em língua inglesa são as definições originais da base de dados da empresa.

Employee includes both internal and external employees (service performers). Unlike externals, internal employee is bound by instructions and obligated to adhere to the company's policies and regulations. A person who contributes or has contributed to the creation of goods or services in the enterprise based on a work contract for services. An employee who does not have an explicit contract for permanent or long-term employment. An individual who performs work for an organization on a non-permanent basis.

a business user who is a third-party employee who is contracted as an external employee.  
 a business user who is a third-party employee participating in a collaboration scenario.  
 a resource element that can operate a device or work without a device.

Employee or external worker.

How could I harmonize the following terms: workforce person, worker, collaboration user, external resource, contingent worker, and employee, considering the definition above. What would be the main term and what would be the variants?

Instruction 4: Consider the following definitions:

A person who contributes or has contributed to the creation of goods and services in the company based on a work contract, an at-will arrangement, or a contract for services. A person who contributes or has contributed to the creation of goods or services for a company.

Employee includes both internal and external employees (service performers). Unlike externals, internal employee is bound by instructions and obligated to adhere to the company's policies and regulations. A person who contributes or has contributed to the creation of goods or services in the enterprise based on a work contract for services. The employee constitutes the main focus of interest for the majority of human resource application components, such as personnel administration, time management, payroll, and so on. All business processes in these application components relate to the employee. An employee who does not have an explicit contract for permanent or long-term employment.

An individual who performs work for an organization on a non-permanent basis. Contingent workers are often known as independent professionals, temporary contract worker, contractors, or consultants.

A business user who is a third-party employee who is contracted as an external employee. For example, external consultants or auditors. Therefore, they share the employment address of the contracted employer for certain period of time.

A business user who is a third-party employee participating in a collaboration scenario, for example vendor-managed inventory.

A collaboration user needs access to a system, for example to maintain stocks.

A resource element that can operate a device or work without a device. Example: forklift driver or picker. An operator on the shop floor that assembles boards, units, or other materials you manufacture.

Employee or external worker.

How could I harmonize the following terms: workforce person, worker, collaboration user, external resource, contingent worker, and employee, considering the definition above. What would be the main term and what would be the variants?

Fonte: Elaborado pela autora.

A estrutura e o conteúdo das instruções em língua inglesa são iguais às geradas em língua portuguesa. A fim de averiguar se há diferença na harmonização entre o par de línguas inglês-português, analisamos as respostas apresentadas pelo modelo GPT-4 32k acerca da instrução 4. Primeiramente, examinaremos a resposta em língua portuguesa (Quadro 15) e, em seguida, a resposta em língua inglesa (Quadro 16).

Quadro 15 - Resposta do GPT-4 32k em português

Instrução 4:

Considere as seguintes definições:

Contrato de trabalho, um acordo sem aviso prévio ou um contrato de serviços. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços para uma empresa.

O funcionário inclui funcionários internos e externos (executores de serviços).

Diferentemente dos externos, o funcionário interno está vinculado a instruções e é obrigado a aderir às políticas e aos regulamentos da empresa. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços na empresa com base em um contrato de trabalho para serviços. O funcionário constitui o principal foco de interesse da maioria dos componentes de aplicação de recursos humanos, como administração de pessoal, gerenciamento de tempo, folha de pagamento e assim por diante. Todos os processos de negócios nesses componentes de aplicação estão relacionados ao funcionário.

Um funcionário que não tem um contrato explícito de emprego permanente ou de longo prazo.

Um indivíduo que trabalha para uma organização de forma não permanente. Os trabalhadores temporários são geralmente conhecidos como profissionais independentes, trabalhadores com contrato temporário, empreiteiros ou consultores.

Um usuário comercial que é um funcionário terceirizado contratado como funcionário externo. Por exemplo, consultores ou auditores externos. Portanto, eles compartilham o endereço de trabalho do empregador contratado por um determinado período de tempo.

Um usuário corporativo que é um funcionário terceirizado que participa de um cenário de colaboração, por exemplo, inventário gerenciado pelo fornecedor.

Um usuário de colaboração precisa de acesso a um sistema, por exemplo, para manter estoques.

Um elemento de recurso que pode operar um dispositivo ou trabalhar sem um dispositivo.

Exemplo: motorista de empilhadeira ou selecionador. um operador no chão de fábrica que monta placas, unidades ou outros materiais que você fabrica.

Funcionário ou trabalhador externo.

Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando as definições acima? Qual seria o termo principal e quais seriam as variantes?

Com base nas definições fornecidas, parece que o termo principal poderia ser "Funcionário". Este termo é amplo o suficiente para abranger todos os tipos de trabalhadores que contribuem para uma empresa, independentemente do tipo de contrato, duração do emprego ou localização.

Aqui estão as variantes:

1. "Empregado" e "Membro do Pessoal": Estes termos são geralmente intercambiáveis com "Funcionário". Eles se referem a qualquer pessoa que trabalhe para uma empresa, independentemente do tipo de contrato.

2. "Colaborador": Este termo é frequentemente usado para se referir a qualquer pessoa que contribua para uma empresa, mas pode ser especialmente relevante para aqueles que não são funcionários tradicionais, como consultores ou freelancers.

3. "Trabalhador Temporário": Este termo se refere a funcionários que têm um contrato de trabalho de curto prazo ou não permanente.

4. "Recurso Externo": Este termo se refere a funcionários que não são internos, como consultores ou auditores externos.

5. "Usuário de Colaboração": Este termo se refere a funcionários terceirizados que precisam

de acesso a um sistema para realizar seu trabalho.

6. "Operário": Este termo se refere a funcionários que operam um dispositivo ou trabalham sem um dispositivo, como um motorista de empilhadeira ou um montador no chão de fábrica.

7. "Trabalhador": Este é um termo genérico que pode se referir a qualquer pessoa que trabalhe para uma empresa, independentemente do tipo de contrato ou posição.

Esses termos podem ser usados de maneira intercambiável em muitos contextos, mas cada um tem suas próprias nuances que podem torná-lo mais apropriado em determinadas situações. A escolha do termo a ser usado dependerá do contexto específico e do público-alvo.

Fonte: OpenAI (2024).

### Quadro 16 - Resposta GPT-4 32k em inglês

Instruction 4: Consider the following definitions:

A person who contributes or has contributed to the creation of goods and services in the company based on a work contract, an at-will arrangement, or a contract for services. A person who contributes or has contributed to the creation of goods or services for a company.

Employee includes both internal and external employees (service performers). Unlike externals, internal employee is bound by instructions and obligated to adhere to the company's policies and regulations. A person who contributes or has contributed to the creation of goods or services in the enterprise based on a work contract for services. The employee constitutes the main focus of interest for the majority of human resource application components, such as personnel administration, time management, payroll, and so on. All business processes in these application components relate to the employee.

An employee who does not have an explicit contract for permanent or long-term employment.

An individual who performs work for an organization on a non-permanent basis. Contingent workers are often known as independent professionals, temporary contract worker, contractors, or consultants.

A business user who is a third-party employee who is contracted as an external employee. For example, external consultants or auditors. Therefore, they share the employment address of the contracted employer for certain period of time.

A business user who is a third-party employee participating in a collaboration scenario, for example vendor-managed inventory.

A collaboration user needs access to a system, for example to maintain stocks.

A resource element that can operate a device or work without a device. Example: forklift driver or picker. An operator on the shop floor that assembles boards, units, or other materials you manufacture.

Employee or external worker.

How could I harmonize the following terms: workforce person, worker, collaboration user, external resource, contingent worker, and employee, considering the definition above. What would be the main term and what would be the variants?

Based on the definitions provided, the main term could be "Workforce Person." This term is broad enough to encompass all the different types of individuals who contribute to a company's operations, whether they are employees, contingent workers, or external resources.

Here's how the other terms could be harmonized as variants:

1. "Worker": This term could be used interchangeably with "Workforce Person." It's a general term that refers to anyone who contributes to the creation of goods or services for a company.



2. "Collaboration User": This term could be a subset of "Workforce Person." It refers to a third-party employee who participates in a collaboration scenario, such as vendor-managed inventory. They need access to a system to perform their tasks.
  3. "External Resource": This term could also be a subset of "Workforce Person." It refers to a business user who is a third-party employee contracted as an external employee, such as external consultants or auditors.
  4. "Contingent Worker": This term is a specific type of "Workforce Person." It refers to an individual who performs work for an organization on a non-permanent basis, such as independent professionals, temporary contract workers, contractors, or consultants.
  5. "Employee": This term is another specific type of "Workforce Person." It refers to a person who contributes to the creation of goods and services in the company based on a work contract, an at-will arrangement, or a contract for services. Employees are bound by the company's policies and regulations.
- In this way, "Workforce Person" becomes the umbrella term, with the other terms serving as more specific variants that fall under this main category.

Fonte: OpenAI (2024).

Ao analisarmos as respostas geradas pelo modelo, percebemos que, em língua portuguesa, o termo principal é funcionário, visto que é um termo mais amplo e que abrange todos os tipos de trabalhadores de uma empresa, independente do contrato empregatício, duração do emprego ou localização. Em seguida, em ordem de frequência e especificidade, o modelo sugere as seguintes variantes: empregado, membro do pessoal, colaborador, trabalhador temporário, recurso externo, usuário de colaboração, operário e trabalhador. Quando analisamos a harmonização em língua inglesa, podemos notar que o termo principal é *Workforce person*, uma vez que ele é um termo amplo e que abrange todos os tipos de funcionário, servindo como um termo guarda-chuva. Com relação às variantes, o modelo sugere os seguintes termos, considerando a ordem do mais específico para o mais genérico: *worker*, *collaboration user*, *external resource*, *contingent worker* e *employee*. Podemos perceber que há diferenças quanto à escolha de termo principal e variantes entre as línguas. Essas diferenças nos levam a crer que tais escolhas não são aleatórias, mas consideram os aspectos sociais, cognitivos, culturais, históricos, legais e institucionais para a harmonização do conjunto de termos em cada língua.

Vale dizer ainda que essa diferença valida a nossa hipótese de que os *frames* podem contribuir para a harmonização de termos e conceitos na medida em que adicionam uma camada extra de informações linguísticas, contextuais e situacionais para a harmonização. Com a elaboração da arquitetura de *frames*, é possível identificar as diferenças linguísticas e estruturais entre as línguas e sociedades e ter uma melhor

compreensão de como harmonizar, já que o comportamento dos termos irá se refletir nessa estrutura. Outro ponto interessante está relacionado à implementação da tarefa de harmonização em empresas multilíngues. Esse experimento indica que é necessário refletir sobre como elaborar a arquitetura de *Knowledge Graph* com *frames* e como aplicar a metodologia semântico-terminológica para harmonização de termos e conceitos em diferentes línguas, uma vez que não se trata de uma mera tradução de toda a abordagem entre um par linguístico, mas deve se considerar todo o contexto, a situação, a conceptualização e o uso da terminologia e aplicar a proposta metodológica para cada língua.

Atentando para essa questão, definimos uma quinta técnica e estratégia de *prompt*, o uso de *Knowledge Graph* no LLM para obtermos uma harmonização mais precisa. Essa decisão é justificada pelo fato de que a harmonização proposta pela ISO não considera os aspectos culturais, sociais e cognitivos na tarefa de harmonizar termos e conceitos. Além disso, nossa proposta semântico-terminológica para a harmonização já prevê a construção de uma representação do conhecimento com *frames*. Portanto, faz-se crucial utilizarmos essa arquitetura tanto na base de dados da empresa como no modelo de linguagem. Para essa instrução, utilizaremos a estrutura do *Knowledge Graph* com *frames* semânticos que desenvolvemos neste trabalho. Ele será inserido no modelo de linguagem para que, ao interpretar a instrução, ele considere o KG e os *frames* para definir o termo principal e as variantes.

A partir da discussão sobre os conceitos e passos metodológicos descritos, evidenciam-se, na Figura 54, todas as etapas previstas para o desenvolvimento da abordagem semântico-terminológica da harmonização de termos e conceitos e sua aplicação em um conjunto de termos. Como podemos observar, apresentamos os doze passos da tarefa para se atingir a harmonização, além de especificarmos os dados que serão apresentados na ficha terminológica.

Figura 54 - Proposta metodológica para a Harmonização de termos e conceitos seguindo a abordagem semântico-terminológica



Fonte: Elaborada pela autora.

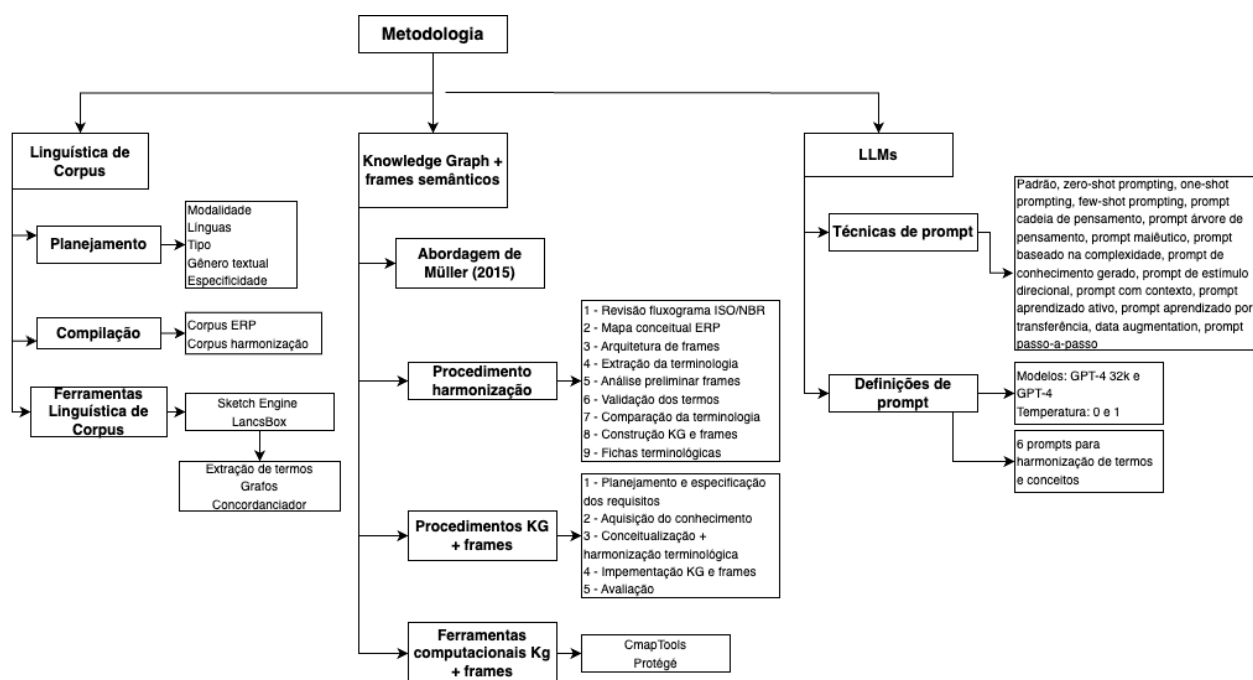
Como indicado na Figura 55, nossa proposta metodológica parte da revisão do fluxograma de harmonização proposto pela ISO. Em seguida, ocupar-nos-emos da compilação e do processamento dos *corpora* da pesquisa para, então, elaborarmos o mapa conceitual do ERP. O quarto passo é realizar a descrição da arquitetura de *frames* semânticos. Posteriormente, extrairemos as listas de termos, que servirão tanto para “popular” nossa representação do conhecimento como para selecionarmos os termos a serem harmonizados. A sexta etapa consiste em analisar preliminarmente a estrutura de *frames* e examinar se está de acordo com a estrutura do *software*. Mais adiante, validaremos a lista de termos e compararemos a terminologia antes do processo de harmonização com a nova listagem de termos. Após isso, construiremos nossa arquitetura de *Knowledge Graph* com *frames*. Com a nova lista de termos já comparada com a terminologia anterior, definiremos os conjuntos de dados a serem harmonizados e definiremos as estratégias de *prompt* a serem aplicadas no modelo para que seja possível

realizar a harmonização de forma automática. O próximo passo é analisar os resultados obtidos no LLM, averiguando se a harmonização está de acordo com o esperado e se o modelo segue a proposta metodológica desenvolvida neste trabalho. Por fim, a última etapa consiste em criar fichas terminológicas para os termos harmonizados, identificando o termo principal, as variantes, classe gramatical, abreviatura/sigla, domínio, equivalentes e exemplos de uso.

Estabelecemos, portanto, procedimentos metodológicos deste trabalho, tratando, inicialmente, da Linguística de *Corpus* e dos *corpora* compilados e utilizados para esta pesquisa, assim como as ferramentas computacionais selecionadas para o processamento e análise do material. Discutimos sobre os *Knowledge Graphs* e os *frames* semânticos como arquitetura a ser implementada tanto experimentalmente numa base de dados como no LLM e as ferramentas utilizadas para a construção dessa estrutura. Por fim, forma abordadas as estratégias e técnicas de *prompt* utilizadas no LLM para a instrução e solicitação de respostas para a tarefa de harmonização terminológica.

Considerando os pontos discutidos neste capítulo, a Figura 55 resume as etapas metodológicas desta tese.

Figura 55 - Síntese do capítulo “Metodologia”



Fonte: Elaborada pela autora.

Por meio da Figura 55, evidencia-se que, inicialmente, tratamos da Linguística de *Corpus*, apresentando o planejamento e a compilação dos *corpora* para este trabalho e as ferramentas computacionais utilizadas para analisar os dados. Depois disso, discorreremos sobre o Knowledge Graph e os *frames* semânticos, apresentando a abordagem metodológica desenvolvida por Müller (2015), adotada para a elaboração da nossa representação do conhecimento. Além disso, também indicamos os passos tomados para o desenvolvimento da harmonização de termos e conceitos de base semântico-terminológica. Na sequência, mostramos os procedimentos estabelecidos para a construção da arquitetura do *Knowledge Graph* com *frames* semânticos e, por fim, listamos ferramentas computacionais para a elaboração da representação do conhecimento. A última etapa da metodologia consistiu em tratar dos *Large Language Models*, tratando das estratégias e técnicas de *prompt* e das definições de *prompt* que selecionamos para este trabalho.

O próximo capítulo, com base nos apontamentos feitos até aqui e nos percursos realizados, traz os resultados alcançados e sua respectiva análise.

## 6 ANÁLISE DE DADOS

Após definirmos as abordagens metodológicas utilizadas neste trabalho, tratando da compilação e do processamento dos *corpora*, das ferramentas selecionadas para a elaboração do *Knowledge Graph* e dos *frames* semânticos e das estratégias de *prompt* que utilizamos em um *Large Language Model* (LLM) (grande modelo de linguagem) para a tarefa de harmonização terminológica, passaremos para a análise dos dados.

Inicialmente, na seção 6.1, descreveremos a construção do *Knowledge Graph* do ERP – *Enterprise Resource Planning* – e dos *frames* semânticos para a harmonização. Posteriormente, na seção 6.2, discutiremos acerca da aplicação da nossa abordagem metodológica para a tarefa de harmonizar três conjuntos de termos, e, finalmente, na seção 6.3, trataremos da harmonização desenvolvida pelo LLM, considerando diferentes estratégias de instrução para essa tarefa e relacionando os dados obtidos do modelo com os resultados encontrados na seção anterior.

### 6.1 Do *Knowledge Graph* aos *Frames*: Representando o conhecimento do ERP

Conforme descrito na metodologia, seguimos a abordagem de Müller (2015) para a construção do *Knowledge Graph* (KG) com *frames* semânticos, a fim de representarmos as informações do ERP. Após as definições metodológicas, tendo em vista que, além da abordagem semântico-terminológica da harmonização de termos e conceitos, também nos propomos a entregar um *Knowledge Graph* do ERP. Nesta seção, trataremos da construção do KG, visto que ele será utilizado como uma das estratégias de *prompt* no LLM (*Large Language Models*), bem como poderá ser implementado experimentalmente na base de dados da empresa. Para isso, como apresentado no capítulo anterior, utilizamos a ferramenta computacional Protégé para a elaboração do *Knowledge Graph*. É importante destacar que o KG foi desenvolvido em língua inglesa, posto que, para disponibilizá-lo, é melhor que a arquitetura seja em inglês.

No que se refere à estruturação do KG, partimos do ERP como uma classe guarda-chuva, uma vez que ele abrange diferentes áreas de uma empresa. Em seguida, como subclasses do ERP, encontramos os seguintes setores empresariais: *Asset*

*Management* (gestão de ativos), *Finance* (finanças), *Human Resources* (recursos humanos), *Manufacturing* (manufatura), *Sourcing & Procurement*, *R&D Engineering* (Engenharia de Pesquisa e Desenvolvimento), *Sales* (vendas), *Service* (serviço), *Supply Chain* (cadeia de suprimentos), *Professional Services* (serviços profissionais), *Other Sectors* (outros setores), *Master Data ERP* e *Country/Region-Specific Functions* (funções específicas para países e regiões). Posteriormente, podemos localizar as demais subclasses de cada setor. Por fim, encontramos as instâncias que descrevem os processos dentro de cada subclasse. Durante o processo de construção do *Knowledge Graph*<sup>57</sup>, atentando para as classes, subclasses, instâncias e relações, o Protégé apresenta um conjunto de métricas relativas à elaboração da representação do conhecimento. No tocante ao nosso KG, encontramos os seguintes dados, ilustrados no Quadro 17:

Quadro 17 - Métricas do *Knowledge Graph*

<b>Métricas</b>	
<b>Classes</b>	601
<b>Axiomas</b>	8.059
<b>Axiomas lógicos</b>	4.463
<b>Axiomas declarativos</b>	3.596
<b>Instâncias</b>	2.987
<b>Relações</b>	8

Fonte: Elaborado pela autora, com base em Musen (2015).

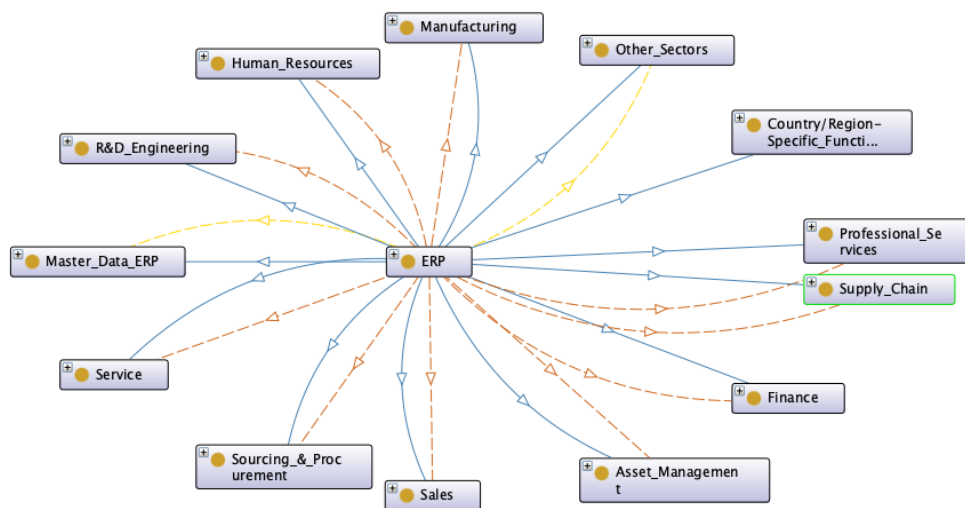
No Quadro 17, podemos observar as métricas do nosso *Knowledge Graph* do ERP. No que se refere às classes, nosso KG é constituído por 601 classes, que representam os conceitos do domínio. Em seguida, encontramos 8059 axiomas, os quais especificam as definições da ontologia a partir de predicados, combinando axiomas lógicos e axiomas declarativos (não-lógicos). Os axiomas lógicos representam estados sempre verdadeiros na constituição e interpretação dos conceitos. Os axiomas declarativos (não-lógicos), por sua vez, são predicados adicionados pelo *software* Protégé e não podem ser vistos na interface do usuário (Musen, 2015). Nossa representação do conhecimento apresenta 2987 instâncias, que remetem aos

<sup>57</sup> Conforme descrito na seção 5.2.2, Ferramentas computacionais para a criação do mapa conceitual e do *Knowledge Graph* + *frames* semânticos, para a elaboração do *Knowledge Graph* do ERP, utilizamos a ferramenta computacional Protégé (Musen, 2015).

elementos/indivíduos da ontologia. Por fim, nosso *Knowledge Graph* é composto por oito relações que representam um tipo de ligação entre os conceitos do domínio.

No que diz respeito à visualização do *Knowledge Graph*, na Figura 56, podemos observar a classe ERP e suas subclasses.

Figura 56 - Classe ERP e suas subclasses

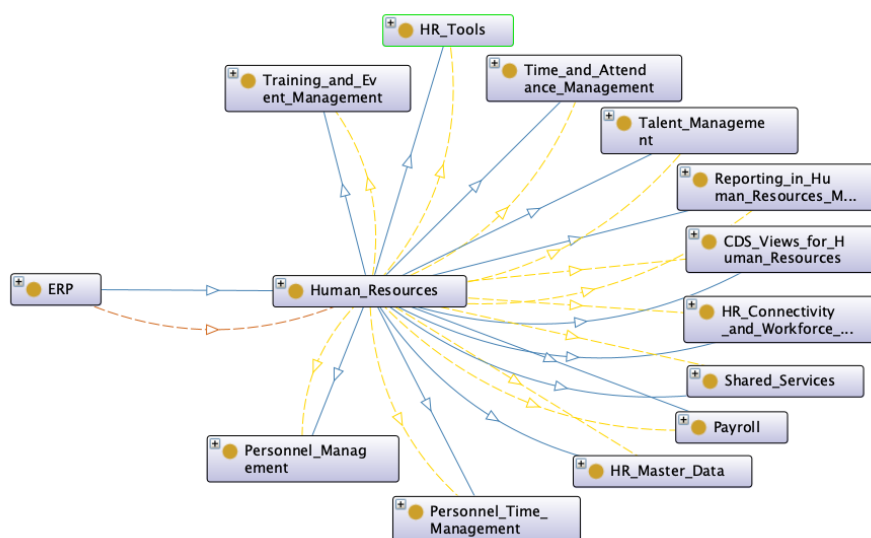


Fonte: Elaborado pela autora.

Com relação às subclasses dos setores do ERP, devido ao tamanho do *Knowledge Graph* e considerando que um dos conjuntos de termos a ser analisado tem relação com o indivíduo que atua em uma empresa, apresentaremos apenas as classes e subclasses do setor de *Human Resources* (recursos humanos) para ilustrar a elaboração do KG e dos *frames*. Tendo em vista que o componente de Recursos Humanos é um dos maiores grafos do nosso KG, apresentando 12 subclasses de recursos do componente e 88 subclasses que representam as funcionalidades dos recursos, para essa exemplificação, traremos a representação de recursos humanos dividida em partes. Inicialmente, podemos vislumbrar o componente RH e seus recursos, conforme a Figura 57.



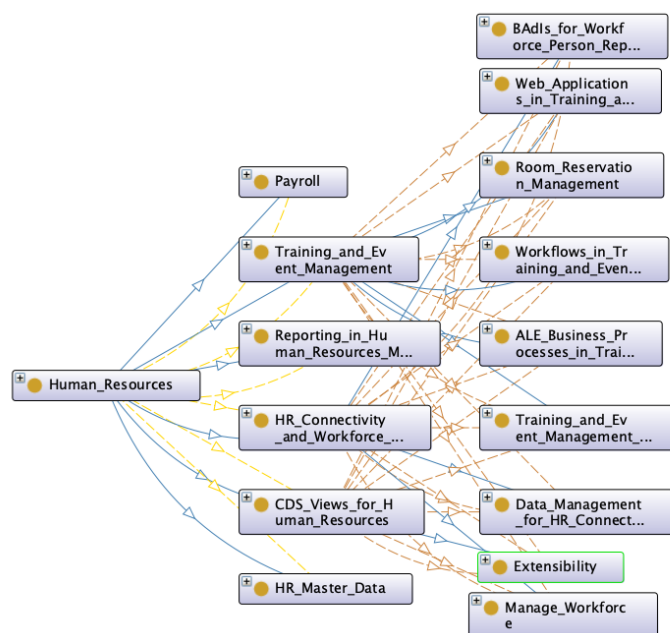
Figura 57 - Classe ERP, subclasse *Human Resource* e componentes



Fonte: Elaborado pela autora.

Conforme ilustra a Figura 58, podemos observar a classe ERP, a subclasse *Human Resources* e seus componentes *Training and Event Management*, *HR Tools*, *Time and Attendance Management*, *Talent Management*, *Reporting in Human Resources Management*, *CDS Views for Human Resources*, *HR Connectivity and Workforce Overview*, *Shared Services*, *Payroll*, *HR Master Data*, *Personnel Time Management* e *Personnel Management*. A fim de visualizarmos as subclasses de cada recurso, apresentaremos o KG de HR em diversas etapas. Na Figura 58, é possível ver o componente, os recursos e as funcionalidades, conforme podemos observar.

Figura 58 - *Knowledge Graph* do setor de *Human Resources* e alguns recursos e funcionalidades

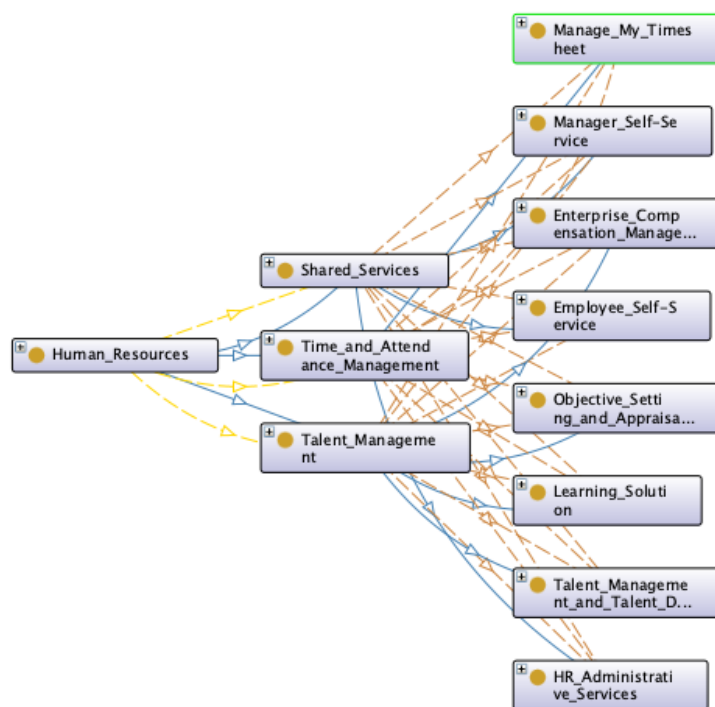


Fonte: Elaborado pela autora.

De acordo com o que ilustra a Figura 58, vemos a estrutura do *Knowledge Graph* do setor de Recursos Humanos. Partindo do primeiro nó (à esquerda da imagem), encontramos o componente *Human Resources* (recursos humanos, doravante HR). Em seguida, temos as subclasses de HR, as quais denominamos de recursos do componente: *Payroll*, *Training and event Management*, *Reporting in Human Resources Management*, *HR Connectivity and Workforce Overview*, *CDS Views for Human Resources* e *HR Master Data*. A última coluna de subclasses é composta pelas funcionalidades dos recursos, como *BADs for Workforce Person Replication*, *Web Application in Training and Event Management*, *Room Reservation Management*, *Workflows in Training and Event Management*, *ALE Business Processes in Training and Event Management*, *Training and Event Management*, *Data Management for HR Connectivity and Workforce Details*, *Extensibility* e *Manage Workforce*.

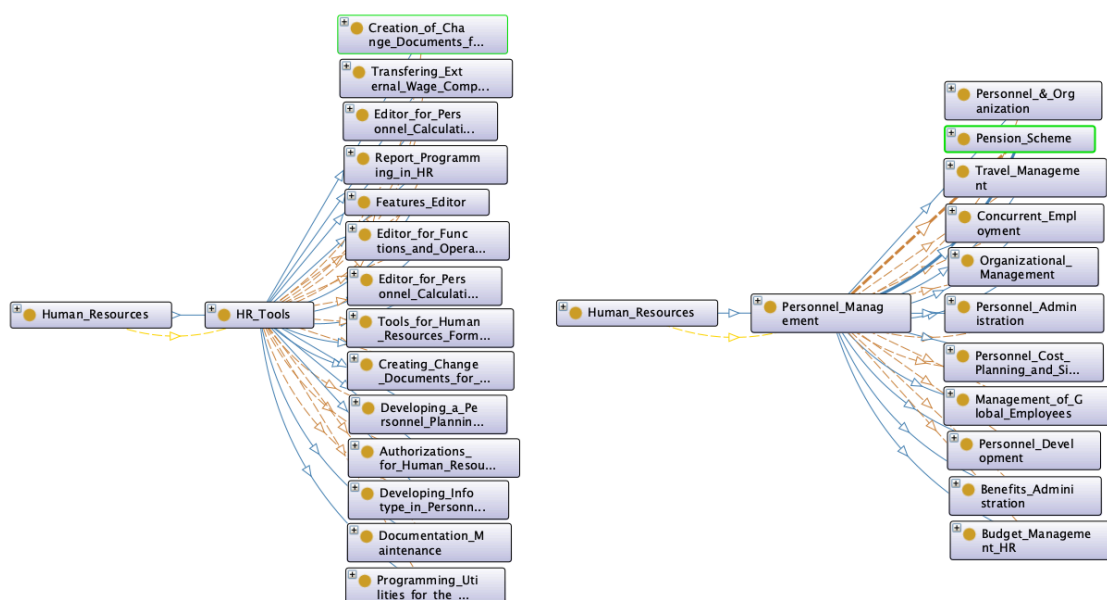
Além desses recursos, encontramos *Shared Services*, *Time and Attendance Management* e *Talent Management*, e os componentes *Manage My Timesheet*, *Manager Self-Service*, *Enterprise Compensation Management*, *Employee Self-Service*, *Objective Setting and Appraisals*, *Learning Solution*, *Talent Management and Talent Development* e *HR Administration Services*. A Figura 59, na sequência, traz a representação desses recursos.

Figura 59 - Componente de Recursos Humanos, recursos e funcionalidades



Fonte: Elaborada pela autora.

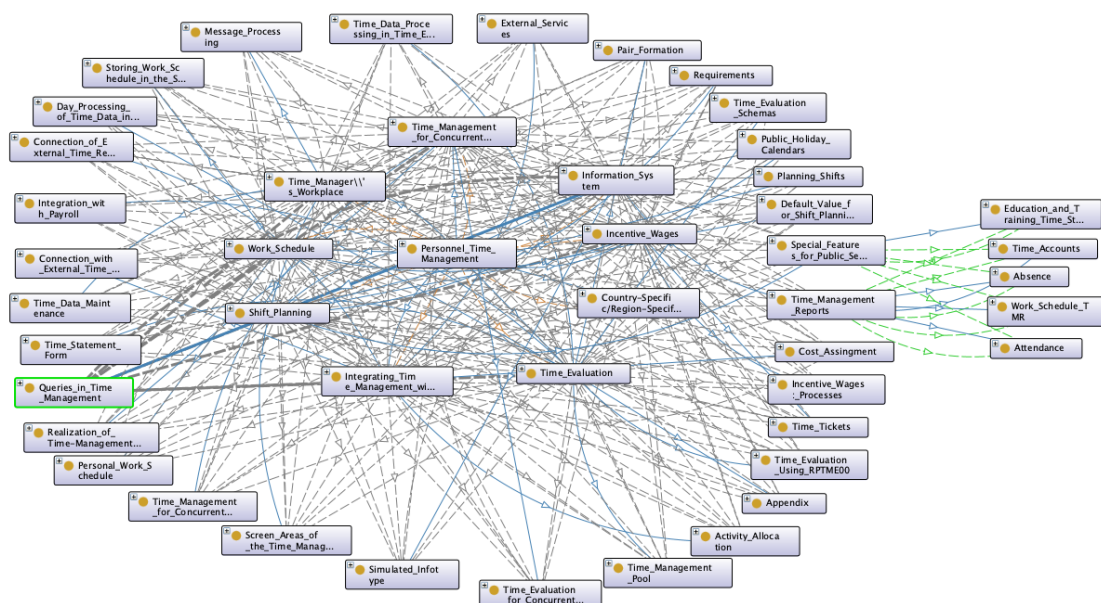
A seguir, podemos examinar os recursos de *HR Tools* e *Personnel Management* e suas funcionalidades, de acordo com o que se apresenta na Figura 60.

Figura 60 - Recursos *HR Tools* e *Personnel Management* e suas funcionalidades

Fonte: Elaborado pela autora.

Com relação ao componente de *Personnel Time Management*, podemos visualizar, por meio da Figura 61, o recurso e suas funcionalidades.

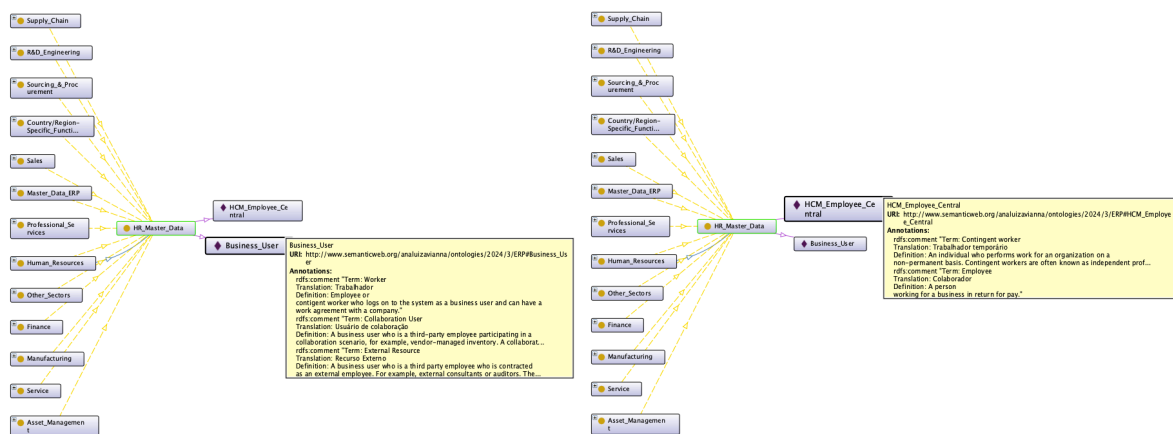
Figura 61 - Recurso *Personnel Time Management* e suas funcionalidades



Fonte: Elaborado pela autora.

Essa parte do *Knowledge Graph* do componente de *Human Resources* mostra a dimensão do grafo. Nela, podemos observar somente o recurso de *Personnel Time Management*, ao centro da imagem, seguido das nove subclasses em seu entorno e, por fim, os processos e funcionalidades que tal recurso apresenta. Esse grafo é um dos maiores recursos do KG, visto que apresenta um alto nível de informações e detalhes do que ocorre nesse recurso.

Além das classes e subclasses, nosso KG também apresenta instâncias e comentários. No que tange aos comentários, nessa parte, adicionamos os termos a serem analisados na harmonização, trazendo a definição terminológica e o equivalente de tradução, como podemos ver na Figura 62.

Figura 62 - Instâncias e comentários do recurso de *HR Tools*

Fonte: Elaborado pela autora.

Concomitantemente ao desenvolvimento da arquitetura do *Knowledge Graph*, elaboraram-se os *frames* semânticos do ERP. Para essa etapa, consideramos a organização da base de dados da FrameNet para a criação dos *frames*. Nesse sentido, baseamo-nos nos pressupostos da base para pensarmos em como estruturar a ordem dos *frames*, sua nomenclatura, as relações e os tipos semânticos. Conforme explicitam Ruppenhofer *et al.* (2010), os *frames* estão situados em um espaço semântico, apresentando relações entre eles e tipos semânticos. Atinente às relações entre *frames*, podemos encontrar os relacionamentos de herança, *subframes*, *causativo\_de*, *incoativo\_de* e *usa*. Em nosso trabalho, podemos observar a presença, principalmente, da relação de *subframes*.

No que se refere aos *subframes*, à medida em que a organização de *frames* é desenvolvida, podemos nos deparar com construções mais complexas, sobretudo na sequência de estados e transições, uma vez que cada *frame*, nesse caso, pode ser descrito separadamente como uma nova estrutura conceptual (Ruppenhofer *et al.*, 2010). Esse tipo de relação pode ser denominado como *subframe*, já que apresenta certa complexidade na sua estrutura, com outros eventos na sua composição e com elementos que podem ser identificados nos demais *frames*. Vale destacar que podemos encontrar *frames* mais complexos dentro da sua estrutura, os quais podem apresentar uma composição rica em unidades lexicais e outras construções.

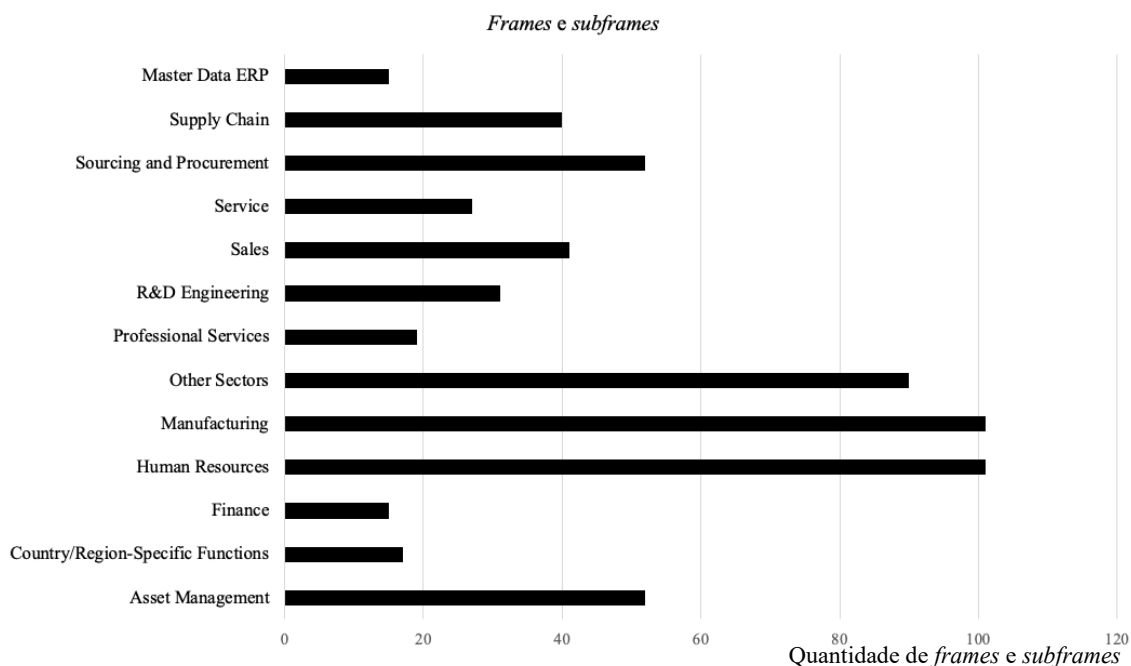
Acerca dos tipos semânticos, Ruppenhofer *et al.* (2010) afirmam que os tipos semânticos na FrameNet servem para registrar informações provenientes da hierarquia dos *frames* e dos elementos de *frame*. Neste trabalho, utilizaremos, essencialmente, o tipo ontológico. Esse tipo semântico representa estruturas que podem ser mapeadas em

uma ontologia, visto que se relaciona com outros *frames* via relação de subtipo. De acordo com Souza (2019), os *frames* ontológicos são mais estáticos, diferente, por exemplo, de *frames* dinâmicos, os quais descrevem eventos e ações. Para a Semântica de *Frames*, os *frames* ontológicos são tipos mais complexos de descrever devido a sua natureza mais eventiva, visto que um *frame*, geralmente, descreve uma ação evocada por um verbo. Mesmo que uma estrutura conceptual não descreva uma ação ou evento, como esperado tradicionalmente, ela ainda é necessária para abarcar um *frame* mais complexo.

Além de esses pontos que buscamos na FrameNet, também nos fundamentamos no formato apresentado pela base de dados para a elaboração dos nossos *frames*. Dessa forma, as estruturas aqui representadas apresentam uma definição, os elementos de *frame*, as unidades lexicais relacionadas ao cenário e as relações.

Tendo em vista toda a criação dos *frames*, considerando os nós do *Knowledge Graph*, identificamos, no total, 601 cenários em nossa arquitetura do ERP. Com relação à distribuição de *frames*, nossa estrutura apresenta um Super *Frame*, 13 *frames* ontológicos, 82 *frames* e 505 *subframes*. Na Figura 63, podemos examinar quantos *frames* e *subframes* cada setor do ERP apresentou na arquitetura do *Knowledge Graph*.

Figura 63 - *Frames e subframes*



Fonte: Elaborado pela autora.

Como podemos observar na Figura 63, o componente '*Master Data ERP*' apresenta 15 *frames* referentes às funcionalidades e aos dados comuns a todos os setores

que compõem o ERP. O setor de *'Supply Chain'* (Cadeia de suprimentos) conta com 40 *frames* relacionados aos recursos e as funcionalidades da gestão de planejamento de estoque e movimentos de mercadorias e produtos. *'Sourcing and Procurement'* tem 52 *frames* para processos de suprimentos operacionais e centrais, bem como o gerenciamento de faturas e integrações. O componente de *'Service'* (serviços) traz 27 *frames* que tratam do ciclo de serviços prestados pela empresa. *'Sales'* (vendas) é composto por 41 *frames* atribuídos a processos e recursos relativos à administração de contratos, dados, preços, ofertas, ordens e faturamentos de vendas. O componente de *'R&D Engineering'* (Engenharia de Pesquisa e Desenvolvimento) indica 31 *frames* direcionados à gestão de portfólios e projetos da empresa. O setor de *'Professional Services'* (serviços profissionais) alcança 19 *frames* que descrevem a gestão de clientes, projetos e recursos. *'Other Sectors'* (outros setores) contabiliza 90 *frames*, incluindo setores como indústria do Petróleo e Gás, Setor Público, Banco e Seguros. Nesse sentido, os *frames* descrevem os recursos e funcionalidades de as áreas. Os componentes de *'Human Resources'* (Recursos Humanos) e de *'Manufacturing'* (Manufatura) apresentam 101 *frames* cada. Tais setores são os que mais apresentam recursos e funcionalidades no ERP devido à complexidade de sua estruturação e à quantidade de dados gerenciados pelas áreas. O setor de *'Finance'* (finanças) indica 15 *frames* que descrevem os recursos e as funcionalidades financeiras disponíveis no *software*. O componente *'Country/Region-Specific Functions'* (funções específicas de países e regiões) tem 17 *frames*, que tratam de funcionalidades diferentes de acordo com as regulamentações de cada país e/ou região. Por fim, *'Asset Management'* conta com 52 *frames* para as estruturas do setor de gestão de ativos de uma empresa, como o planejamento de recursos e gestão de manutenção.

No que diz respeito ao início do arranjo da estrutura conceptual, compreendemos que o ERP pode ser considerado como um Super *Frame*, uma vez que ele abrange todos os demais *frames*. Nesse caso, o ERP seria um *frame* guarda-chuva, posto que ele não recebe muitas instâncias nem *subframes*, servindo como base para a estruturação dos outros setores. Tendo em vista a importância da camada de informações advinda dos *frames* semânticos para a representação do conhecimento e para a compreensão dos termos, a seguir, apresentaremos o Super *Frame* do ERP e os *frames* e *subframes* da área de Recursos Humanos para exemplificarmos as descrições elaboradas para cada setor. Os *frames* desenvolvidos apresentam o nome do *frame*, uma definição, elementos de *frame*, unidades lexicais, tipos de *relações* e relações entre *frames*. Na definição, os

elementos de *frame* estão destacados com realce em diferentes cores para fins de anotação de etiquetas, seguindo os modelos da FrameNet (Ruppenhofer *et al.*, 2010) No Quadro 18, podemos examinar a descrição do Super *Frame* ERP.

Quadro 18 - Super *Frame* ERP

Frame: Enterprise_Resource_Planning (ERP)
<b>Definição:</b> Solução que auxilia na gestão de toda uma empresa, abrangendo diferentes <b>componentes</b> , isto é, os setores empresariais, como Manufacturing, Human_Resources, Sales, R&D_Engineering, Services, Finance, Asset_Management, Source_&_Procurement, Other_Services, Country/Region-Specific_Functions, Supply_Chain e Professional_Services. Tal programa permite o acesso fácil, integrado e confiável aos dados corporativos, além de fornecer diagnósticos relevantes sobre cada setor, contribuindo para as tomadas de decisões e futuros passos necessários para reduzir custos e aumentar a produção da empresa.
<b>Elementos de <i>frame</i>:</b> <b>Componentes:</b> Setores de uma empresa que fazem parte do ERP. Cada área do <i>software</i> apresenta diferentes ferramentas que contribuem para a automação e os processos daquela esfera.
<b>Unidades lexicais:</b> Manufacturing, Human Resources, Sales, R&D Engineering, Services, Finance, Asset Management, Source & Procurement, Other Sectors, Country/Region-Specific Functions, Supply Chain, Professional Services.
<b>Tipos de relações:</b> <i>Is_composed_by</i> e <i>Includes</i>
<b>Relações entre <i>frames</i>:</b> Manufacturing, Human_Resources, Sales, R&D_Engineering, Services, Finance, Asset_Management, Source_&_Procurement, Other_Services, Country/Region-Specific_Functions, Supply_Chain, Professional_Services.

Fonte: Elaborado pela autora.

No Super *Frame* ERP, como podemos observar no Quadro 18, é apresentado o elemento de *frame* ‘componentes’ (em realce amarelo). Pensando na arquitetura do *Knowledge Graph* e na sequência de *frames*, os componentes representam os setores que o compõem no ERP, i.e., manufatura, recursos humanos, finanças, dentre outras áreas. Com relação aos *frames* dos componentes, cada setor do ERP pode ser entendido como um *frame* ontológico. Como dito anteriormente, o *frame* ontológico, conforme apresentado por Ruppenhofer *et al.* (2010) e Souza (2019), é mais estático e, geralmente, não recebe muitas unidades lexicais. No nosso caso, o *frame* ontológico de cada setor do ERP exerce a função de estruturar toda a arquitetura de cada setor e conectar os *frames* específicos de cada área com o Super *Frame* ERP. No Quadro 19,



podemos ver a descrição do *frame* ontológico Human Resources (Recursos Humanos).

Quadro 19 - *Frame* ontológico Human\_Resources

Frame: Human_Resources
Definição: Componente que representa a área focada na gestão de pessoas da organização. A área de Recursos Humanos se ocupa do recrutamento e seleção de novos talentos, engajamento entre empregados, empregador e demais setores empresariais, treinamento e desenvolvimento pessoal, benefícios, planos de carreira e remuneração. O componente apresenta diferentes <b>recursos</b> para gestão organizacional da empresa.
Elementos de <i>frame</i> : <b>Recursos</b> : Cada componente apresenta diferentes recursos, que auxiliam no gerenciamento do setor. Em <i>Human Resources</i> (recursos humanos), os recursos contribuem para a gestão de pessoas, treinamento e desenvolvimento do pessoal, gerenciamento de remuneração, horários e eventos. Nessa área encontramos os seguintes recursos disponíveis: CDS_Views_for_Human_Resources, HR_Connectivity_and_Workforce_Overview, HR_Master_Data, HR_Tools, Payroll, Personnel_Management, Personnel_Time_Management, Reporting_in_Human_Resources_Management, Shared_Services, Talent_Management, Time_and_Attendance_Management, Training_and_Event_Management.
Unidades lexicais: (O <i>frame</i> Human_Resources não apresenta unidades lexicais, visto que ele é um <i>frame</i> ontológico)
Tipos de relações: <i>Manages</i>
Relações entre <i>frames</i> : ERP, Manufacturing, Sales, R&D_Engineering, Services, Finance, Asset_Management, Source_&_Procurement, Other_Services, Country/Region-Specific_Functions, Supply_Chain, Professional_Services.

Fonte: Elaborado pela autora.

É importante salientar que o *frame* Human\_Resources, por ser considerado como um *frame* ontológico, não apresenta unidades lexicais, apenas a definição, elementos de *frame* e tipos de relações. Seguindo a estrutura de *frames*, após o *frame* ontológico, encontramos os *frames* de cada área do ERP. A título de exemplo, no Quadro 20, podemos ver a descrição do *frame* Personnel\_Management.

Quadro 20 - *Frame* Personnel\_Management

Frame: Personnel_Management
Definição: Personnel Management é um recurso disponível no setor de Recursos Humanos. Tal recurso envolve diferentes <b>funcionalidade</b> que auxiliam na aplicação de estratégias relativas aos empregados e gestão da empresa, contribuindo para o desenvolvimento da instituição.
Elementos de <i>frame</i> : <b>Funcionalidade</b> : As funcionalidades são ferramentas disponíveis para auxiliar no

gerenciamento de recursos. Em Manufatura, as funcionalidades permitem que se administre todos os empregados, novas contratações, desenvolvimento de talentos, pagamentos e eventos.

Unidades lexicais: Hire, new hire, web hire, employee, applicant administration, benefits administration, compensation administration, pension fund administration, personnel administration, system specialist personnel administration, work center and job administration, age-related benefit, car benefit, core benefit, deferred benefit, housing benefit, insurance benefit, interest benefit, invalidity benefit, leaving benefit, projected benefit, retirement benefit, standard benefit, surviving dependent's benefit, taxable housing benefit, compassionate leave, compensatory leave, leave, concurrent employment, desired employment, employee with local terms of employment, employment, global employment, multiple employment, offer to employment, off-cycle workbench for concurrent employment, period of employment, place of employment, prohibition of employment, terms and conditions of employment, time evaluation for concurrent employment, applicable large employer, employer, previous employer, assign, assignment, country/component assignment, employee organizational employment, length of position assignment, length of talent assignment, main personnel assignment, matrix assignment, organization assignment, organization reassignment, period of global assignment, personnel assignment, personnel cost assignment, position assignment, staff assignment, staff reassignment, successor assignment, talent group assignment, user assignment, vacancy assignment, further training candidate, grant date, high-date, hiring date, hold date, imaginary date, initial entry date, learned date, leaving date, low-date, notice date, notification date, performance date, pre-hire start date, publication date, retirement date, RP update, start date, succession candidate, task date, update, valid from date, valid to date, validity end date, validity start date, vested date, budget assigned per person/position, cash position, chief position, collective position, direct position, employee information by position, employee basic pay position, filled position, indirect position, key position, length of time in position, nomination context for key position, occupied position, position, staff position, unoccupied position, vacant position, flex fund, health insurance fund, military compensation fund, model fund, budget management, compensation management, employee relations management, enterprise compensation management, human capital management, human resources management, organizational management, performance management, personnel budget plan management, position management, state management, talent relationship management, internal data, long term incentives exercising data, number of applicants with master data, payroll data, plan data, planned compensation data, planning data, planning-relevant data, recruiting data, salary survey data, schedule data, transfer data, work center data, confidential data, communication data, composite market data, cost-relevant data, flexible employee data, general benefits data, job, preferred job, previous job, survey job, total compensation, pregnancy compensation, total cash compensation, planned compensation, period of continued pay, projected pay, take home pay, lump-sum payment, overdue payment, value of payment, Benefits, compensation management, personnel management, pension fund, user management, data protection, communication channel security, communication destinations, job posting, position, position management, employee central, fire, effective date, system of record.

Tipos de relações: *Has*

Relações entre *frames*: CDS Views for Human Resources,

HR_Connectivity_and_Workforce_Overview, HR_Master_Data, HR_Tools, Payroll, Personnel_Management, Personnel_Time_Management, Reporting_in_Human_Resources_Management, Shared_Services, Talent_Management, Time_and_Attendance_Management, Training_and_Event_Management.
---

Fonte: Elaborado pela autora.

Como podemos observar, o *frame* `Personnel_Management` apresenta uma definição, seus elementos de *frame*, suas unidades lexicais, seus tipos de relação e suas relações entre *frames*. Como elementos de *frame*, encontramos ‘funcionalidade’, isto é, ferramentas que auxiliam na gestão do componente. Após o *frame* de `Personnel_Management`, podemos vislumbrar os seguintes *subframes*: `Benefits_Administration`, `Budget_Management_HR`, `Concurrent_Employment`, `Managment_of_Global_Employee`, `Organizational_Management`, `Pension_Scheme`, `Personnel_&_Organization`, `Personnel_Administration`, `Personnel_Cost_Planning_and_Simulation`, `Personnel_Development` e `Travel_Management`. No Quadro 21, podemos explorar o *subframe* `Personnel_Administration`.

Quadro 21 - *Subframe* `Personnel_Administration`

Frame: <code>Personnel_Administration</code>
Definição: <code>Personnel_Administration</code> é uma das funcionalidades disponíveis no setor de Recursos Humanos. Essa funcionalidade possibilita gerir todas as tarefas de administração de pessoas, como transferir os dados de empregados contratados, adicionar automaticamente novos empregados na estrutura organizacional da empresa, transferir a compensação do funcionário diretamente para a funcionalidade de pagamento e acionar o desenvolvimento de pessoas.
Elementos de <i>frame</i> : (O <i>frame</i> <code>Personnel_Administration</code> não apresenta elementos de <i>frame</i> , visto que ele é o final do processo de gestão de pessoas na área).
Unidades lexicais: <code>Absence condition day</code> , <code>absence condition gap width</code> , <code>absence exception rule</code> , <code>absence exception subtype</code> , <code>academic level</code> , <code>access log</code> , <code>accessories</code> , <code>accommodation</code> , <code>accounting process</code> , <code>acquisition of qualification</code> , <code>action menu</code> , <code>action reason</code> , <code>action type</code> , <code>additional personal data</code> , <code>address change</code> , <code>address change form</code> , <code>address to contact in case of emergency</code> , <code>adjusted benefit value</code> , <code>adjustment accident</code> , <code>adjustment interval</code> , <code>adjustment run/payroll</code> , <code>administration of employee data</code> , <code>administrative act</code> , <code>administrative event</code> , <code>administrative status</code> , <code>affects remuneration</code> , <code>age</code> , <code>age for pay scale jump</code> , <code>age structure</code> , <code>agreement area</code> , <code>amf number</code> , <code>annual tax reporting</code> , <code>annual tax statement</code> , <code>appraisal</code> , <code>archiving subject</code> , <code>assign</code> , <code>assignment to wage level</code> , <code>ats data collector</code> , <code>average headcount FTE</code> , <code>average headcount – 12 months</code> , <code>average headcount – 4 quarters</code> , <code>bank details</code> , <code>basic pay</code> , <code>basic salary</code> , <code>basis interval</code> , <code>bearer of secrets</code> , <code>benefit scheme</code> , <code>benefit subscheme</code> , <code>benefit value</code> , <code>birth type</code> , <code>birthday list</code> , <code>blueprint framework</code> , <code>blueprint table</code> , <code>branch of study</code> , <code>business category</code> , <code>business journey</code> , <code>calculation of time specifications</code> , <code>calculation process</code> , <code>car code</code> , <code>car compensation</code> , <code>car loan</code> , <code>care of</code> , <code>certificate</code> , <code>certificate type</code> , <code>challenge</code> ,

challenge group, challenge type, challenged person, change in staffing level, change of country, change of locality, child number, city, collective agreement area, collective agreement provision, collective agreement type, collision group, communication, company car, company instructions, company pensioner, company structure, compensation debt, completed training activity, compulsory pension insurance, concatenated field, conditional amount, context authorization check, context authorization main switch, context authorization object, context problems, context solution, contract element, conversion class, conversion rule, corporate function, corporate number, corporate reduction, cost compensation, cost distribution, country of birth, creation of infotype log, data collector for annual tax statement, data privacy, date, date monitoring, date of leaving, date specifications, date status, date/deadline, deadline, decision operation, defaults for pay scale reclassification, degree of challenge, department text, diploma, directory of severely challenged persons, dismissal, dissolution contract, division, economic effects, education and training, educational establishment type, educational institute, elementary school, emergency contact, employee basic pay item, employee basic pay position, employee count, employee data, employee family/related person, employee information, employee information by type and subtype, employee information by position, employee internal control data, employee list, employee master data, employee master document, employee object on loan, employee one-off payment, employee organizational assignment, employee profile, employee recurring payment/deductions, employee search, employee status, employment, employment period calculation, enterprise structure, entry into civil service, entry rate, examination area, examination result, executive employee, exemption for challenged persons, expatriate send/receive indicator, external bank transfers, extraordinary notice of dismissal, fast data entry, field group, fixed-term contract, fixed-term labor contract, follow-up time, follow-up examination, functional area, gratuity, health examination, hiring, home country, hour of travel, inactive work relationship, illness type, income limit, increase, industrial worker, inhouse agreement, instant dismissal, instruction, insurance amount, insurance cover, insurance policy, insurance tariff, insurer, interest pattern, internal and external recruitment, internal medical service, interrupted work relationship, labor contract, language of communication, learning, leave, leave entitled days, leave travel allowance, leavers, leaving, leaving notification, leaving rate, left company, length of pay scale membership, line manager, list of maternity data, long-term document, manager generic personnel administration, mandatory training activity, marital status, marking for organizational reassignment, mass letter, medical check-up program, medical examination, military compensation fund, military exercise, monitoring of tasks, multipurpose wage type, municipality key, name affix, non-leave entitled day, number of employees, number of leavers, number of personnel actions, office employee, office of federal contract compliance programs, organizational assignment, other taxable benefit, other/previous employers, part owner, part ownership, part period, pay scale area, pay scale reclassification, pay scale structure, payment in lieu of notice, payment interval, payment model, payroll status, pensions reserves group, percentage breakdown of employees, percentage breakdown of leavers, performance appraisal, permanent work relationship, permanent incapacity to work, permanently unable to work, person group, person selection period, person with security clearance, personal data, personal identification number, personnel area, personnel control record, personnel subarea, personnel structure, place of birth, plant section, power of attorney, previous employer, previous insurance, probation, probation period, probationary period,

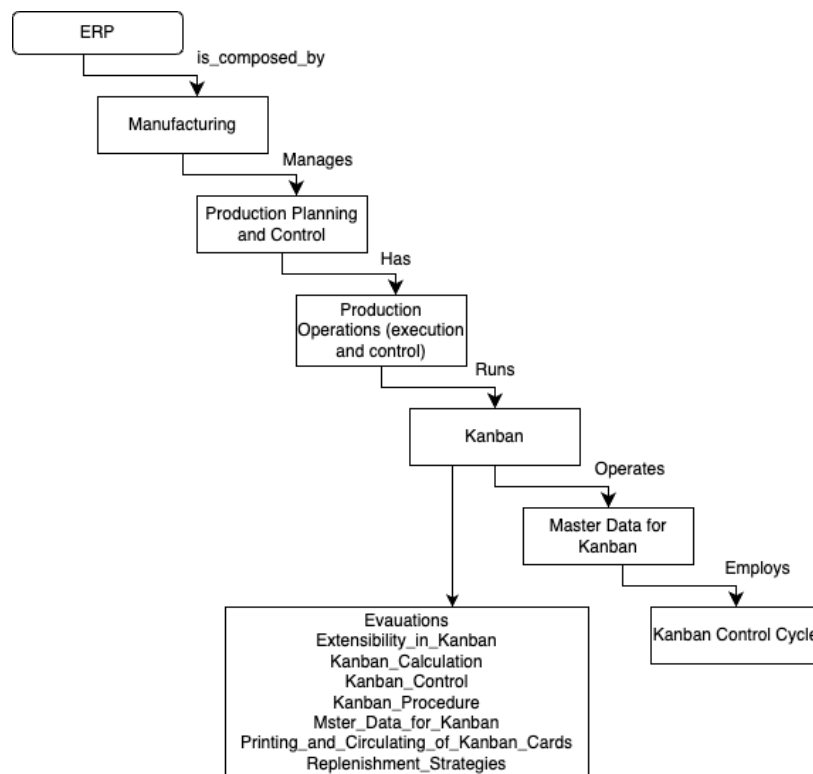
qualifications, reassignment to wage group, reassignment to wage level, recurring advance, recurring payments/deductions, rehiring, relationship to child, relocation, rent allowance, rounding rule, salaried employee, school graduate, school leaving certificate, security class, security officer, severe challenge, severely challenged person, shift entitle, short-term document, sideline job, specialist, staffing status, start of work, status feature, subject of examination, subsidiary company, talent details, task, task date, task record, task reminder, task type, task-oriented labor contract, tax adjustment amount, tax adjustment percent, tax exemption for children, tax form, tax identification number, tax type, taxable, taxable benefit, taxable cost benefit, taxable housing benefit, taxable interest benefit, team grid, temporary address, temporary work relationship, test producers, time specification type, title, trailer record, training period, transfer and promotion, transportation allowance, transportation expenses, travel costs, trip, union department, unlimited contract, unlimited labor contract, upcoming training activity, update log, vacation debt, valuation model, valuation rule, vehicle search list, wage type structure, work contract, work history, work related medical examination, worker's council.

Tipos de relações: *Runs*

Relações entre *frames*: Benefits\_Administration, Budget\_Management\_HR, Concurrent\_Employment, Management\_of\_Global\_Employees, Organizational\_Management, Pension\_Scheme, Personnel\_&\_Organization, Personnel\_Cost\_Planning\_and\_Simulation, Personnel\_Development e Travel\_Management

Fonte: Elaborado pela autora.

No *subframe* Personnel\_Administration, podemos ver uma grande quantidade de unidades lexicais, uma vez que tal funcionalidade tem por objetivo gerir todas as pessoas que trabalham na empresa/instituição/organização, tratando de pagamentos, gestão de pessoas e talentos, desenvolvimento de pessoas, fundos, vagas de emprego, admissão e demissão de empregados, contratos de trabalho, informações pessoais e empregatícios de empregados, dentre outros dados. Vale frisar que esse *subframe*, Personnel\_Administration, não apresenta outros elementos de *frame*, pois ele é o final do processo de gestão de pessoas nessa área, sendo, então, populado por instâncias. Em outros *frames*, como, por exemplo, em Manufatura, há uma continuidade de *subframes* até chegarmos às instâncias, como podemos observar na Figura 64.

Figura 64 - Parte do *Knowledge Graph* de *Manufacturing*

Fonte: Elaborado pela autora.

Por meio da Figura 64, podemos observar uma parte do *Knowledge Graph* de *Manufacturing*, vislumbrando os nós e as relações entre eles. No que se refere aos nós (*frames* e *subframes*) do KG, partimos do ERP, que é composto pelos (relação: *is\_composed\_by*) componentes, aqui representado pelo *Manufacturing* (Manufatura). Depois, *Manufacturing* gerencia (relação: *manages*) *Production Planning and Control* (Planejamento e Controle de Produção), que tem (relação: *has*) *Production Operations (execution and control)* (Operações de Produção (Execução e Controle)). Em seguida, Operações de Produção executa (relação: *runs*) a metodologia Kanban, a qual opera (relação: *operates*) *Master data for Kanban*. De Kanban, podemos observar que esse sistema apresenta os seguintes dados: *Evaluations*, *Extensibility in Kanban*, *Kanban Calculation*, *Kanban Control*, *Master Data for Kanban*, *Printing and Circulating of Kanban Cards* e *Replenishment Strategies*. Por último, temos *Master Data for Kanban* que utiliza (relação: *employs*) *Kanban Control Cycle*.

Após apresentarmos a construção dos *frames* e a arquitetura do *Knowledge Graph* com *frames* semânticos, na próxima seção, trataremos dos conjuntos de termos a serem harmonizados. Nesta etapa, abordaremos a aplicação da nossa proposta de metodologia semântico-terminológica para a harmonização de termos e conceitos.

## 6.2 A metodologia semântico-terminológica para Harmonização de termos e conceitos

Como já discutido na seção 3.3 Harmonização de conceitos e termos: definições e práticas, a abordagem para harmonização de termos e conceitos proposta pela ISO não considera os aspectos sociais, culturais e cognitivos na sua aplicação. Para suprir essa lacuna, nossa metodologia segue uma abordagem semântico-terminológica, considerando não apenas o termo *per se*, mas também a estrutura semântica e conceptual decorrente dos *frames*. Essa base cognitiva permite que compreendamos melhor a organização do domínio ERP, bem como a dos demais setores do *software*, a fim de fornecer mais informações linguísticas e contextuais para a seleção do termo principal e suas variantes.

Com o objetivo de testarmos a funcionalidade e a eficiência da nossa abordagem metodológica, selecionamos três conjuntos de termos para a análise. No capítulo anterior, discorremos sobre o uso dos *corpora* tanto para a elaboração do KG e dos *frames* como para a seleção dos termos a serem harmonizados. Para a etapa de seleção de termos da harmonização, inicialmente, analisamos qualitativamente os termos dos *corpora* e criamos uma lista com a terminologia que ocorria em mais de um setor do ERP. Após essa listagem, buscamos pelos termos que apresentavam uma sobreposição e eram similares. À medida que analisamos o *corpus* em português e em inglês, coletamos termos em ambas as línguas, ainda que, neste trabalho, aplicaremos apenas a harmonização nos termos em língua portuguesa. Posteriormente, buscamos, na base de termos da empresa, as definições dos termos, suas traduções (via coleta bidirecional, por meio da qual buscamos tanto as traduções do inglês para o português como do português para o inglês) e em quais componentes a terminologia está inserida. Essa etapa de identificação dos componentes é fundamental para a construção dos *frames*, uma vez que auxilia na organização das unidades lexicais dentro do ERP e das estruturas conceptuais, assim como contribui para a nossa abordagem, indicando as informações necessárias para compreender o termo e seu contexto.

Nesta tese, após a análise dos *corpora* e identificação da terminologia a ser harmonizada, selecionamos os seguintes termos para realizarmos os experimentos com a nossa abordagem, conforme indicam os quadros apresentados na sequência.

Quadro 22 - Conjunto de termos 1

<b>Termos em inglês</b>	<b>Termos em português</b>
<b>Employee</b>	Empregado Colaborador Funcionário
<b>Contingent Worker</b>	Trabalhador temporário
<b>External Resource</b>	Recurso externo
<b>Collaboration User</b>	Usuário de colaboração
<b>Worker</b>	Funcionário Operário Trabalhador
<b>Workforce Person</b>	Membro do pessoal

Fonte: Elaborado pela autora.

Quadro 23 - Conjunto de termos 2

<b>Termo em inglês</b>	<b>Termo em português</b>
<b>Quotation</b>	Oferta Cotação
<b>Bid</b>	Lance

Fonte: Elaborado pela autora.

Quadro 24 - Conjunto de termos 3

<b>Termo em inglês</b>	<b>Termos em português</b>
<b>Item set</b>	Conjunto de itens
<b>Lot</b>	Lote Pacote

Fonte: Elaborado pela autora.

Considerando os três conjuntos apresentados, aplicaremos nossa proposta de metodologia semântico-terminológica para a harmonização de termos e conceitos, selecionando o termo principal e as variantes e, posteriormente, desenvolvendo uma ficha terminológica para cada conjunto. Para esse trabalho, separamos cada grupo em subseções para realizarmos as devidas análises. Vale salientar que, neste trabalho, nossa aplicação da metodologia e a análise dos termos serão feitas apenas em língua portuguesa, ainda que tenhamos descrito nos quadros 22, 23 e 24 os termos em língua inglesa.

### 6.2.1 Conjunto de termos 1: Empregado

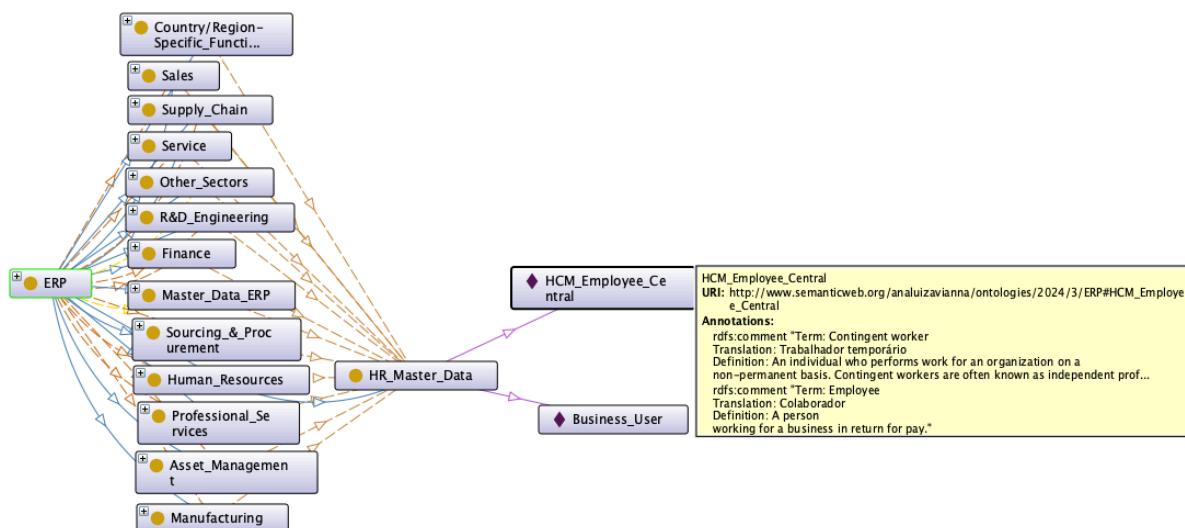
O primeiro conjunto de termos a ser analisado é composto pela terminologia relativa à pessoa que trabalha para uma instituição ou organização. Tal grupo terminológico é composto pelas seguintes unidades lexicais: empregado, colaborador,



funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal. Para esta análise, consideramos as definições dos termos disponíveis na base de dados da empresa, os componentes nos quais tais unidades lexicais estão inseridas, os *frames*, definições de dicionários e demais materiais informativos.

No que se refere às definições dos termos, a unidade ‘colaborador’ é definida na base terminológica como alguém que trabalha para uma empresa em troca de pagamento. De acordo com o dicionário Caldas Aulete (2009)<sup>58</sup>, colaborador pode ser definido como alguém “que colabora, que ajuda pessoa ou grupo num trabalho qualquer. Que ajuda outra pessoa a exercer a sua função”. No que tange aos *frames* e aos componentes, ‘colaborador’ está ligado ao componente de Recursos Humanos, no *subframe* de HR\_Master\_Data, conforme mostra a Figura 65.

Figura 65 - *Subframe* HR\_Master\_Data com anotação do termo 'colaborador'



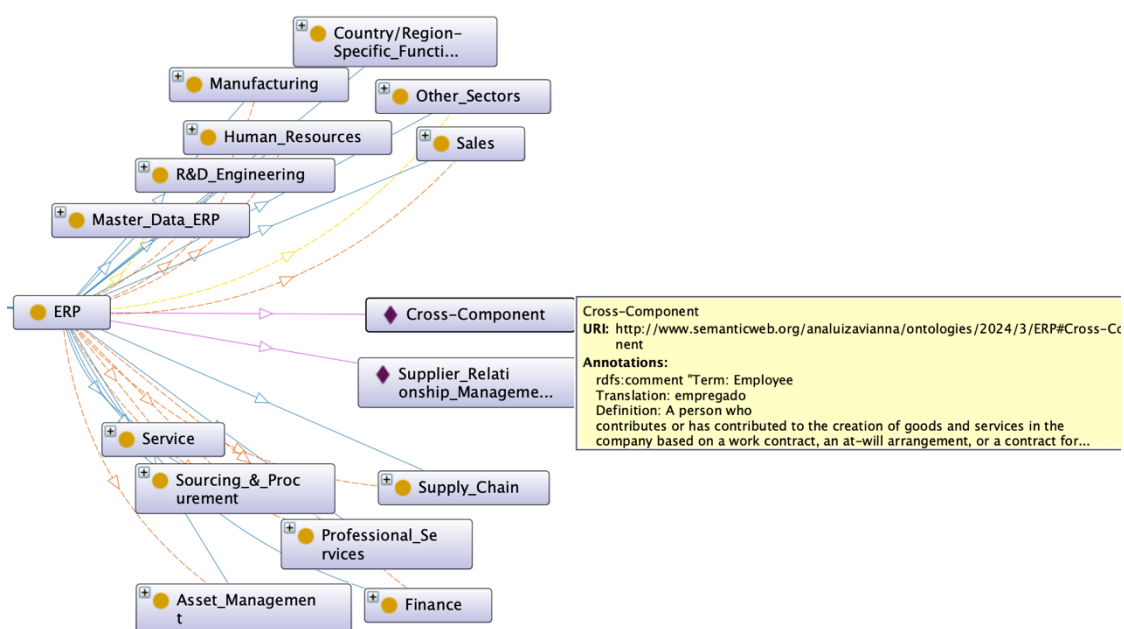
Fonte: Elaborado pela autora.

O termo ‘empregado’, na base terminológica, apresenta três definições: (i) alguém que contribui ou contribuiu para a criação de bens e serviços em uma empresa a partir de um contrato de trabalho, um acordo entre as partes ou um contrato de prestação de serviço; (ii) um link que leva à tela de Funcionário para exibir funções e informações sobre a pessoa; e (iii) alguém que contribui ou contribuiu para a criação de bens e serviços em uma empresa a partir de um contrato de trabalho ou contrato de prestação de serviço. O empregado é o principal foco de interesse para a maioria das

<sup>58</sup> Definição da unidade lexical ‘colaborador’ retirada do Dicionário Caldas Aulete online. Disponível em: <https://aulete.com.br/colaborador> Acesso em: 1 de maio 2024.

funcionalidades em Recursos Humanos, como, por exemplo, administração do pessoal, gestão de tempo, folha de pagamento etc. Segundo o dicionário Caldas Aulete (2009)<sup>59</sup>, empregado pode ser definido como alguém “que se empregou. Que foi colocado em uso, aplicado. Que foi admitido em emprego. Aquele que tem função ou emprego remunerado em firma, empresa, etc.”. No que diz respeito à localização, ‘empregado’ se encontra no Super *Frame* ERP, visto que é um termo utilizado em diversos componentes. Além disso, ‘empregado’ também pode ser localizado no componente de Recursos Humanos, *subframe* *Personnel\_Management*, como podemos observar nas figuras 66, 67 e 68.

Figura 66 - Termo 'empregado' localizado no Super *Frame* ERP, instância *Cross-Component*



Fonte: Elaborado pela autora.

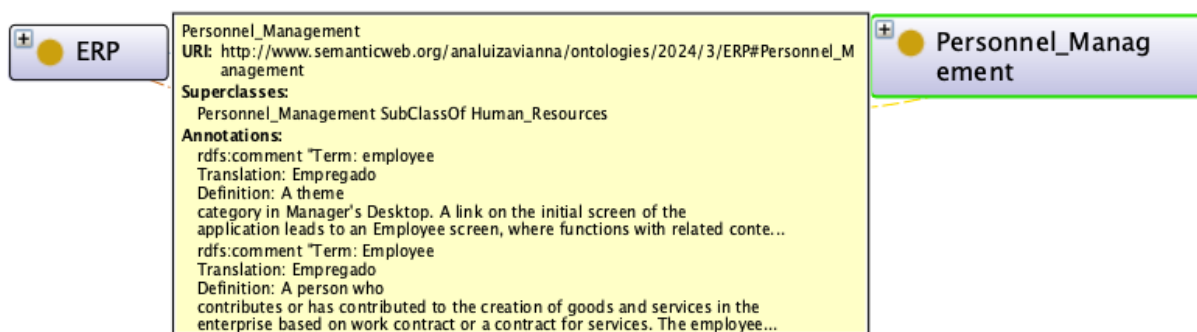
Figura 67 - *Human Resources* e *subframe* *Personnel\_Management*



Fonte: Musen (2015).

<sup>59</sup> EMPREGADO. *In.*: Dicionário Caldas Aulete Online. Disponível em: <https://aulete.com.br/empregado> Acesso em: 1 maio 2024.

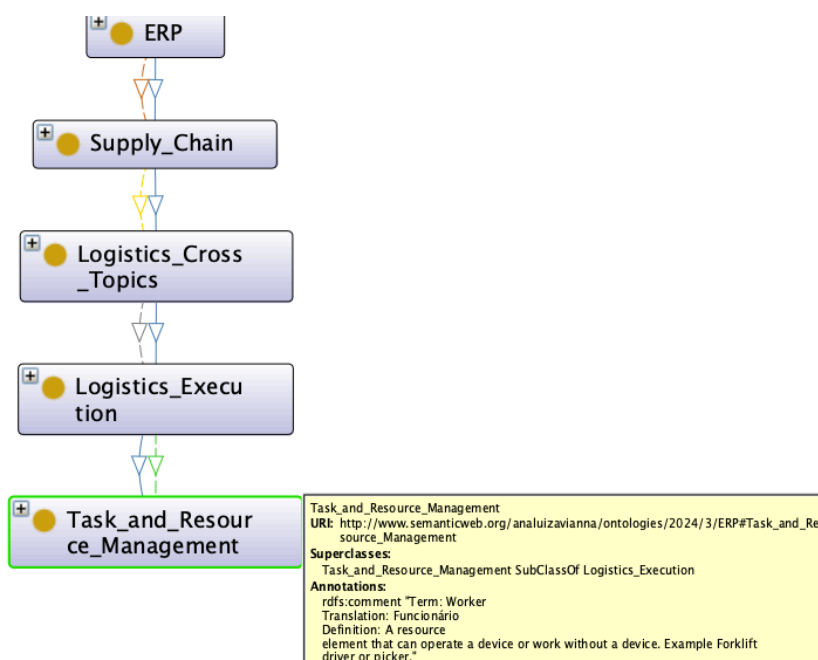
Figura 68 - Anotação Personnel\_Management



Fonte: Elaborado pela autora.

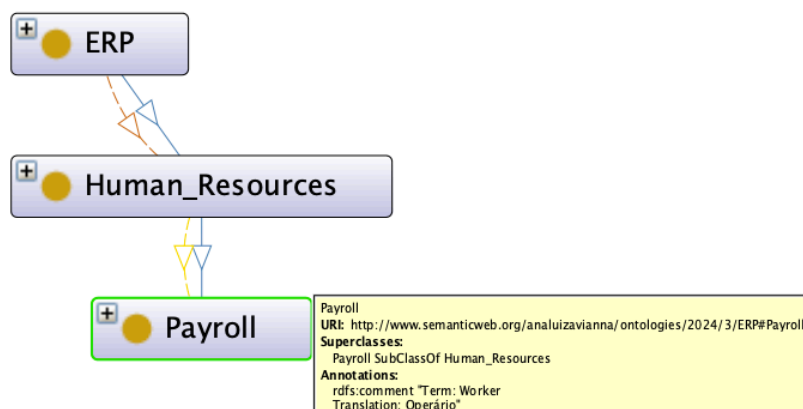
Com relação ao termo ‘funcionário’, encontramos duas definições: (i) um usuário de uma solução; e (ii) alguém que pode operar um dispositivo ou trabalhar sem um dispositivo, por exemplo, motorista ou selecionador de empilhadeira. Conforme descrito no dicionário Caldas Aulete (2009)<sup>60</sup>, funcionário se refere à “pessoa que desempenha função em estabelecimento comercial, empresa etc. Aquele que é empregado ou contratado numa firma, empresa, instituição, repartição”. No caso dos componentes e *frames*, o termo ‘funcionário’ está inserido no componente de *Supply Chain*, mais especificamente no *subframe* `Task_and_Resource_Management`, como mostra a Figura 69.

<sup>60</sup>FUNCIONÁRIO. In.: Dicionário Caldas Aulete Online. Disponível em: <https://aulete.com.br/funcion%C3%A1rio> Acesso em: 1 maio 2024.

Figura 69 - *Subframe* Task\_and\_Resource\_Management

Fonte: Elaborado pela autora.

O termo ‘operário’ é definido, na base terminológica, como alguém que opera no chão de fábrica, que monta placas, unidades ou outros materiais fabricados pela empresa. Para o dicionário Caldas Aulete (2009), operário se refere a “qualquer pessoa que exerce uma ocupação manual mediante o pagamento de salário. Pessoa encarregada do trabalho mecânico ou manual em indústrias e fábricas. Qualquer trabalhador, artífice ou artesão.”. Acerca do termo no *Knowledge Graph*, a unidade está localizada no componente de Recursos Humanos, no *subframe* Payroll (Figura 70).

Figura 70 - *Subframe* Payroll

Fonte: Elaborado pela autora.

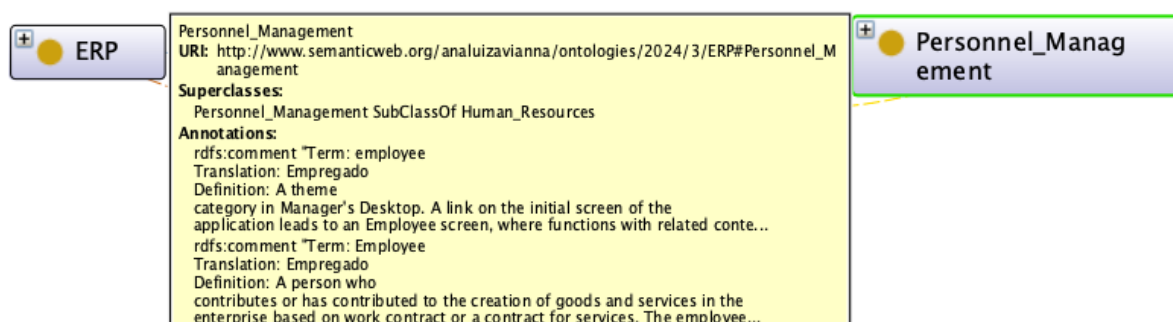
A unidade ‘recurso externo’ apresenta as seguintes definições na base terminológica: (i) alguém que presta serviços comprometidos para completar uma tarefa de um projeto; e (ii) um usuário que é um funcionário terceirizado contratado como funcionário externo, os consultores ou auditores externos, por exemplo. Eles compartilham o endereço de trabalho do empregador por um determinado período de tempo. Tal unidade não apresentou ocorrência nos dicionários. Atinente à localização do termo, encontramos duas ocorrências: a primeira, no componente de Recursos Humanos, em específico no *subframe* de *Personnel\_Administration*, e a segunda, na instância *Business\_User*, localizada no *subframe* de *HR\_Master\_Data* (Recursos Humanos), como podemos observar nas figuras 71, 72 e 73.

Figura 71 - *Human Resources* e *Personnel\_Management*



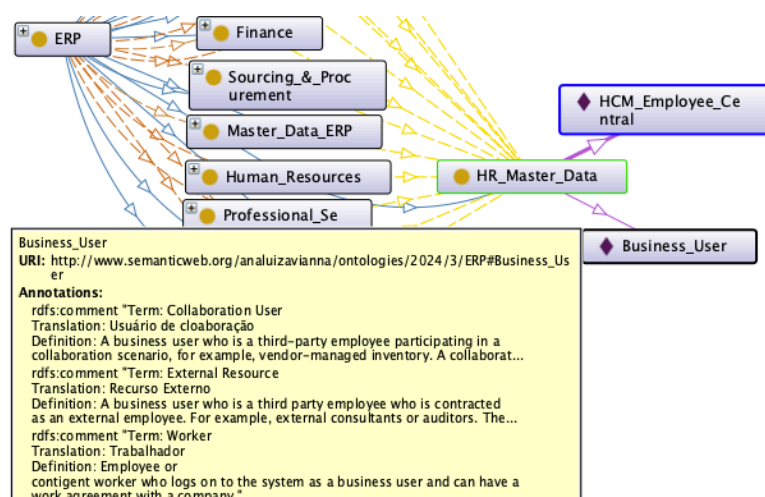
Fonte: Elaborado pela autora.

Figura 72 - Anotação *Personnel\_Management*



Fonte: Elaborado pela autora.

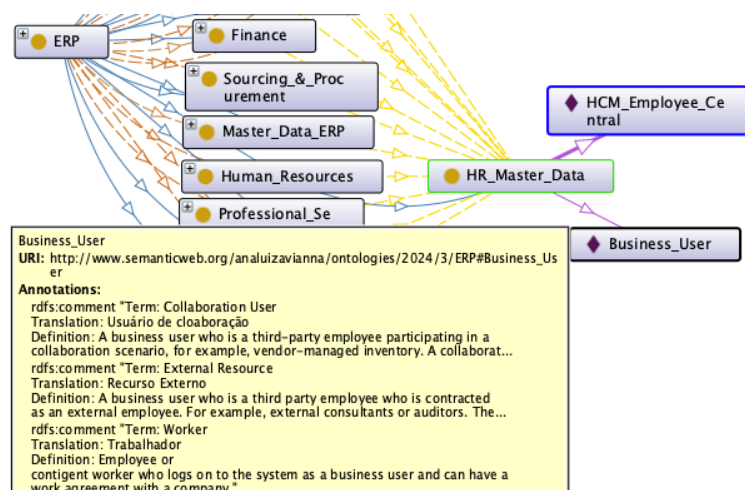
Figura 73 - Anotação instância Business\_User



Fonte: Elaborado pela autora.

Para o termo ‘trabalhador’, a base terminológica define a unidade como funcionário ou trabalhador temporário que utiliza o sistema como usuário e pode ter um contrato de trabalho com a empresa. De acordo com o dicionário Caldas Aulete (2009)<sup>61</sup>, trabalhador é definido como alguém “que trabalha (indivíduo trabalhador). Que se dedica com esmero, afincos etc. à execução de tarefas. Pessoa que trabalha; empregado; operário”. Em nosso *Knowledge Graph*, tal termo está localizado no componente de Recursos Humanos, no *subframe* de HR\_Master\_Data, na instância de *Business\_User* (Figura 74).

Figura 74 - Anotação instância Business\_User

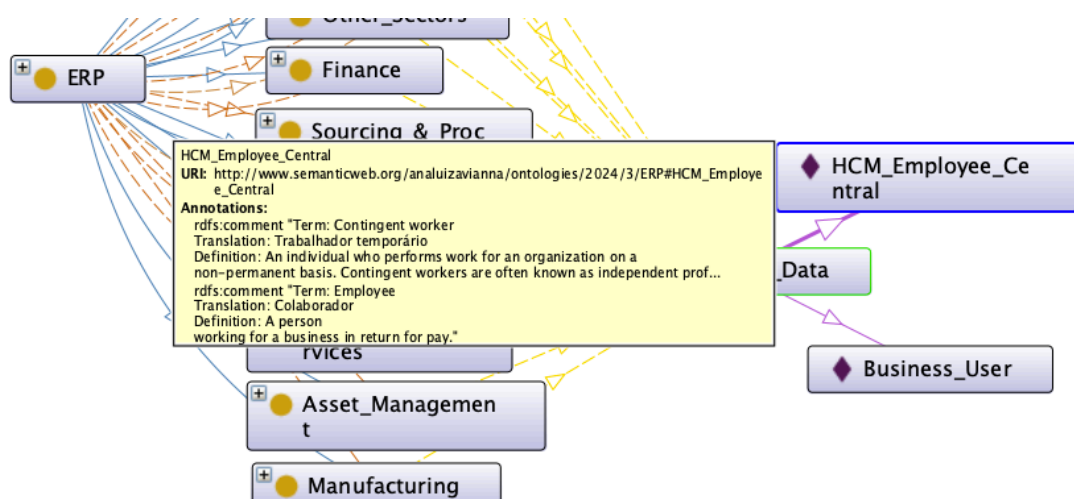


Fonte: Elaborado pela autora.

<sup>61</sup> TRABALHADOR. In.: Dicionário Caldas Aulete Online. Disponível em: <https://aulete.com.br/trabalhador> Acesso em: 1 maio 2024.

‘Trabalhador temporário’, conforme a base terminológica, pode ser definido como (i) um empregado que não tem um contrato explícito de vínculo empregatício permanente e (ii) alguém que realiza um trabalho para uma empresa de forma não permanente. Os trabalhadores contingentes são, geralmente, conhecidos como profissionais independentes, trabalhadores temporários, empreiteiros ou consultores. Com relação às definições dicionarísticas acerca do termo ‘trabalhador temporário’, realizamos as buscas pelo termo ‘trabalhador’ e não encontramos a composição ‘trabalhador temporário’, apenas “trabalhador autônomo, trabalhador braçal, trabalhador migratório e trabalhador rural”. Em termos de localização no KG, ele se encontra no componente de Recursos Humanos, no *subframe* de HR\_Master\_Data, na instância *HCM\_Employee\_Central* (Figura 75).

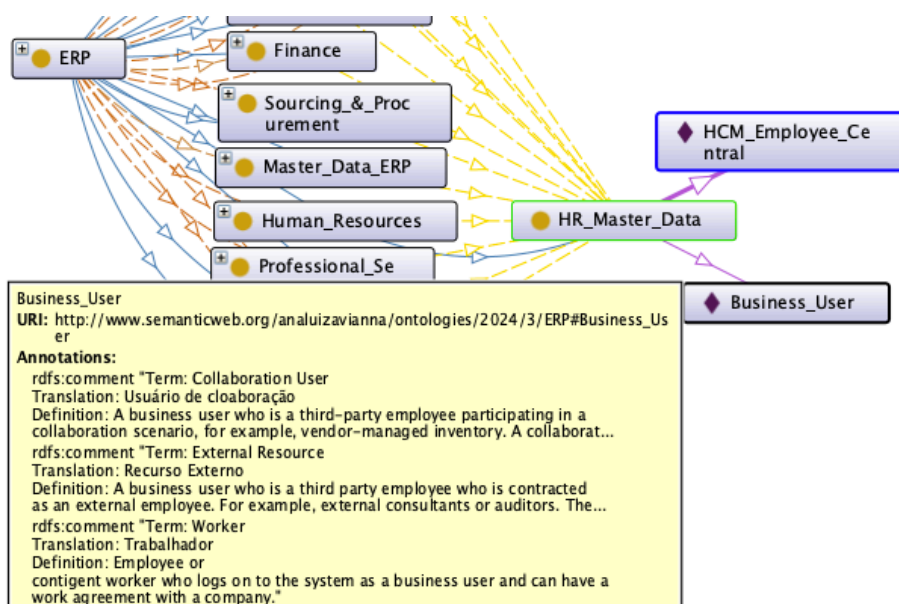
Figura 75 - Anotação instância HCM\_Employee\_Central



Fonte: Elaborado pela autora.

Por fim, o último termo, ‘usuário de colaboração’, é definido na base terminológica como um usuário que é um funcionário terceirizado e que participa de um cenário de colaboração, em inventário gerenciado pelo fornecedor, por exemplo. Um usuário de colaboração necessita de um acesso ao sistema para desempenhar suas tarefas. Este termo não apresentou ocorrências nas buscas em dicionários. Tal unidade está localizada no componente de Recursos Humanos, no *subframe* de HR\_Master\_Data, na instância *Business\_User* (Figura 76).

Figura 76 - Anotação instância Business\_User



Fonte: Elaborado pela autora.

Considerando todas as definições tanto da base terminológica como dos dicionários consultados e as informações do *Knowledge Graph* e dos *frames*, selecionamos o termo ‘empregado’ como termo principal, uma vez que ele faz parte do *Super Frame* ERP e é utilizado em diversos componentes do *software*. Além disso, observamos que, para definir os outros termos do conjunto, usou-se como referência à unidade ‘empregado’. Outro ponto de destaque para essa escolha diz respeito às leis trabalhistas. De acordo com a legislação brasileira, no Decreto-Lei N° 5.452, de 1 de maio de 1943, Artigo 3, “considera-se empregado toda pessoa física que prestar serviços de natureza não eventual a empregador, sob a dependência deste e mediante salário” (Brasil, 1943). Sob o viés legal, observa-se no Decreto-Lei que o termo ‘empregador’ é utilizado para se referir à pessoa que emprega. Nesse sentido, é mais adequado selecionar, pensando na relação de trabalho e na estruturação dos *frames*, ‘empregado’ como termo principal desse conjunto de unidades.

Com relação aos termos variantes, percebemos que, pelas definições, o termo ‘colaborador’, ainda que pudesse ser visto como um possível termo principal, não apresenta tantas especificações e detalhamentos como o termo ‘empregado’. Além disso, as definições designam ‘colaborador’ a partir do referente ‘empregado’. Vale salientar, ainda, que o termo ‘colaborador’ tem em sua definição a relação com alguém que colabora o que implica, de certo modo, um trabalho (não) remunerado, tal qual uma colaboração. Pensando nisso, ‘colaborador’ não seria tão representativo a ponto de ser o



termo principal desse conjunto de termos. A mesma situação pode ser vista com os termos ‘funcionário’ e ‘trabalhador’. Funcionário poderia abarcar todos os tipos de empregados. Entretanto, para definir ‘funcionário’, também se utilizou como referente ‘empregado’. Além disso, tendo em vista que tal termo, em nosso KG, está localizado no componente *Supply Chain* e apresenta uma definição que se refere a um indivíduo com uma função específica como motorista, tal unidade não poderia ser selecionada como termo principal. No caso do termo ‘trabalhador’, tanto as definições da base de termos como dos dicionários usam como referente ‘funcionário’ e ‘empregado’ para tratar do ‘trabalhador’.

Em suas definições, os termos ‘operário’, ‘trabalhador temporário’, ‘recurso externo’ e ‘usuário de colaboração’ apresentam exemplos bem específicos de tipos de trabalhos desenvolvidos por empregados e suas funções na empresa. A título de exemplo, as definições de ‘operário’ indicam um tipo de trabalho próprio, como o de alguém que opera no chão de fábrica. Embora o operário tenha os mesmos direitos e apresente um contrato de trabalho, tal termo designa um papel particular dentro da empresa. O conjunto dos termos ‘trabalhador temporário’, ‘recurso externo’ e ‘usuário de colaboração’ está mais próximo de alguém que é terceirizado ou consultor e não apresenta um contrato permanente nem direto com a empresa empregadora. Logo, não poderíamos selecionar tais termos como principais, uma vez que se referem a um papel específico que determinada pessoa desempenha em uma empresa, organização ou instituição e que não tem funções básicas de um empregado.

A partir da harmonização feita manualmente, desenvolvemos uma ficha terminológica, na qual seguimos a estrutura descrita na seção 5.2.1 Criação do *Knowledge Graph* e *frames* semânticos e Harmonização de termos e conceitos. Assim, são evidenciadas as etapas da tarefa de harmonização e as informações relativas à criação das fichas terminológicas. A seguir, podemos visualizar a ficha terminológica do termo ‘empregado’ (Quadro 25).

Quadro 25 - Ficha terminológica do termo ‘empregado’

<b>Termo:</b> Empregado
<b>Entrada:</b> Empregado
<b>Classe gramatical:</b> sm.
<b>Abreviatura/sigla:</b> Emp.
<b>Domínio:</b> Recursos Humanos
<b>Definição:</b> Indivíduo que contribui ou contribuiu para a criação de bens ou prestação de serviços em uma empresa (empregador) sob um contrato de trabalho. O empregado abrange o pessoal interno, que trabalha sob contrato permanente, como o pessoal externo, que presta serviços terceirizados ou consultorias.
<b>Fonte da definição:</b> Desenvolvido pela autora
<b>Contexto:</b> Recursos Humanos
<b>Variantes:</b> Funcionário, colaborador, membro do pessoal, trabalhador, trabalhador temporário, recurso externo, usuário de colaboração, operário.
<b>Equivalentes:</b> Employee, contingent worker, external Resource, worker, collaboration user.

Fonte: Elaborado pela autora.

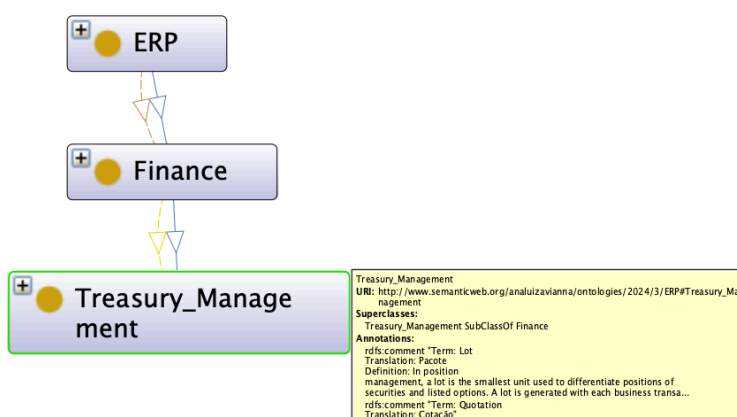
No Quadro 25, podemos observar a ficha terminológica para o conjunto de termos composto por ‘empregado’, ‘colaborador’, ‘funcionário’, ‘operário’, ‘trabalhador’, ‘trabalhador temporário’, ‘usuário de colaboração’ e ‘recurso externo’. É interessante destacar que, com relação às variantes, não há uma ordem específica. Se implementada a harmonização na base de dados da empresa, ao digitar uma das variantes no *software*, por exemplo, a unidade lexical já seria substituída automaticamente pelo termo principal.

### 6.2.2 Conjunto de termos 2: Cotação

O grupo de termos composto por ‘cotação’, ‘lance’ e ‘oferta’ apresenta contextos diferentes, ainda que haja uma relação muito próxima entre eles e, de certo modo, há uma sequência entre a terminologia. Pensando nisso, buscamos as definições na base terminológica, bem como em dicionários especializados. Além disso, procuramos situar a terminologia no *Knowledge Graph* e nos *frames* para compreendermos melhor as unidades e selecionarmos o termo principal e as variantes.

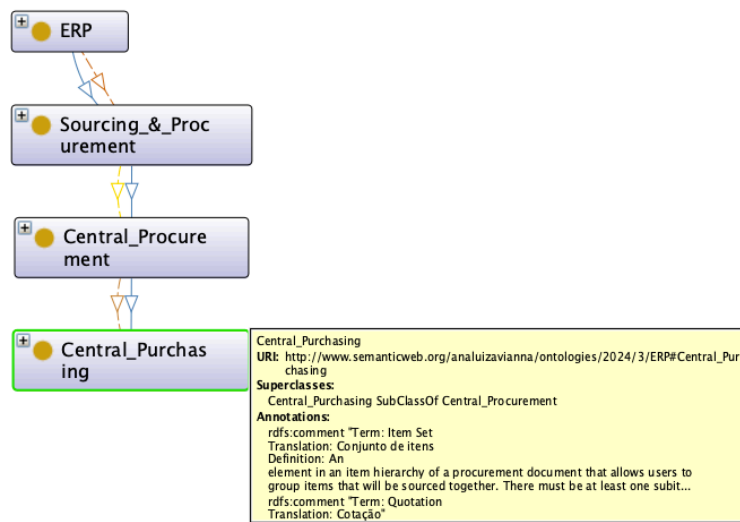
O termo ‘cotação’ apresenta quatro definições na base terminológica. A primeira se refere a como uma oferta é efetuada por um fornecedor a uma organização/empresa para o fornecimento de materiais e para a prestação de serviços sob determinados termos e condições. A segunda definição diz respeito a uma resposta de um determinado fornecedor a uma solicitação de cotação (*request for quotation*). A solicitação de cotação é um documento no qual o fornecedor apresenta as quantidades concretas de produtos que pode fornecer e os prazos de entrega que consegue cumprir. A terceira definição está relacionada a um preço público cotado para um determinado produto em um determinado momento. A última definição está relacionada a uma oferta detalhada de preços, datas de entrega e condições de pagamentos emitidas pelo vendedor. No que tange aos componentes e *frames*, tal unidade está localizada nos componentes de *Finance*, *subframe* *Treasury\_Management*, *Supply Chain*, na instância *Supply\_Chain\_Cockpit*, *Sourcing & Procurement*, *subframes* *Central\_Purchasing* e *Sourcing\_and\_Contract\_Management*, *Master\_Data\_ERP*, *subframe* *Mass\_Maintenance\_of\_Products*, e *Other Sectors*, mais especificamente em *Oil\_and\_Gas*, conforme as figuras 77, 78, 79, 80, 81, 82. De acordo com o Dicionário do Marcílio (2024), cotação está relacionada à cotação de preços, de ações e ao pregão.

Figura 77 - *Subframe* *Treasury\_Management*



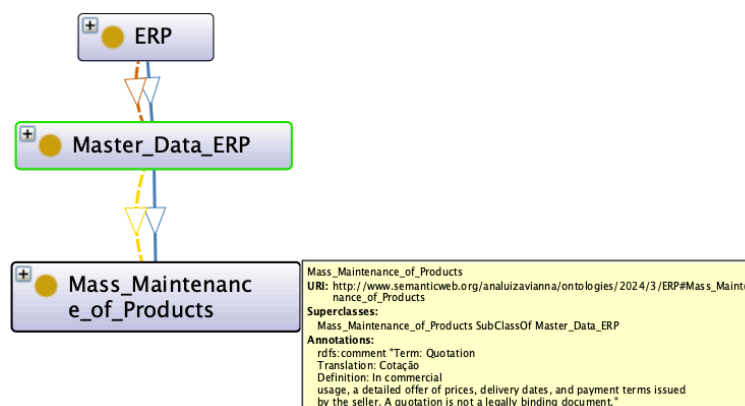
Fonte: Elaborado pela autora.

Figura 78 - *Subframe* Central\_Purchasing



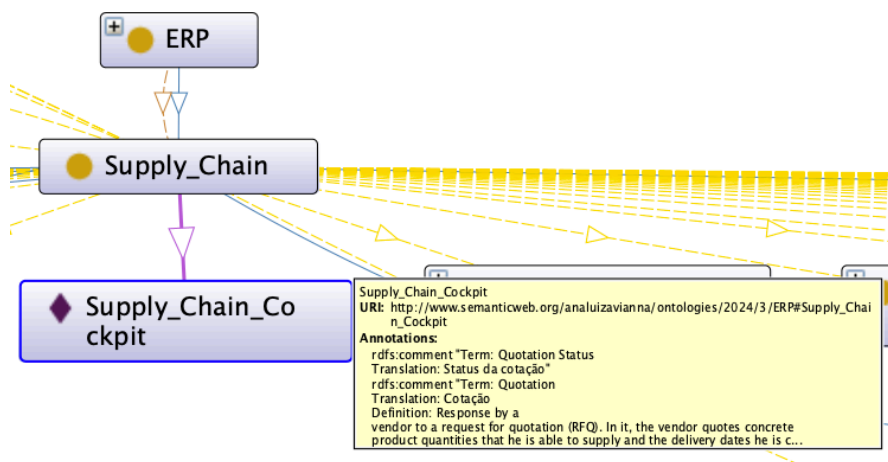
Fonte: Elaborado pela autora.

Figura 79 - *Subframe* Mass\_Maintenance\_of\_Products



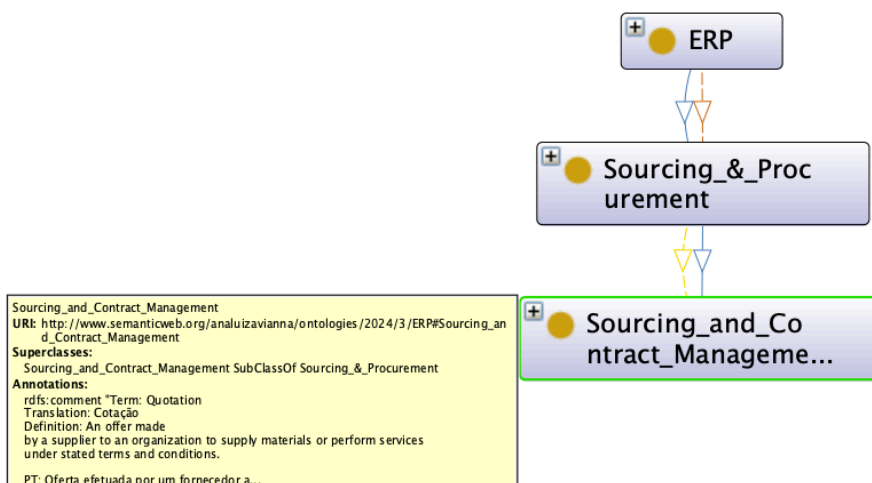
Fonte: Elaborado pela autora.

Figura 80 - Anotação instância *Supply\_Chain\_Cockpit*



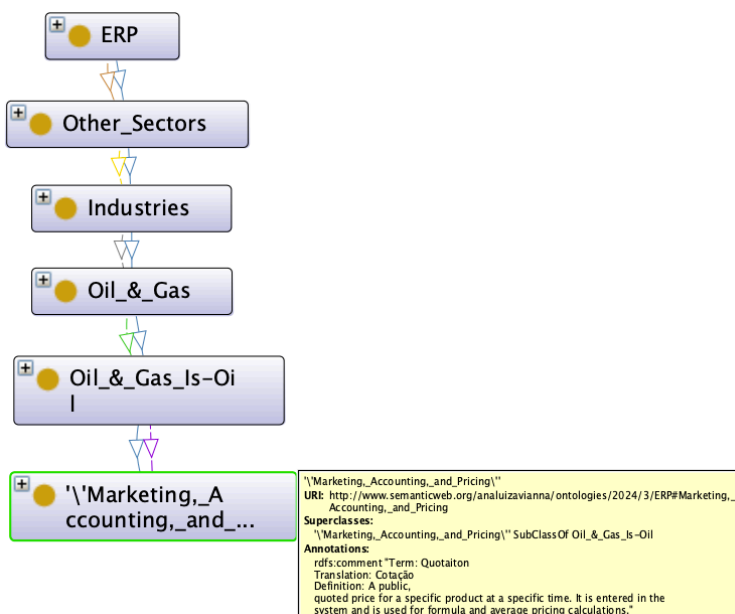
Fonte: Elaborado pela autora.

Figura 81 - *Subframe* Sourcing\_and\_Contract\_Management



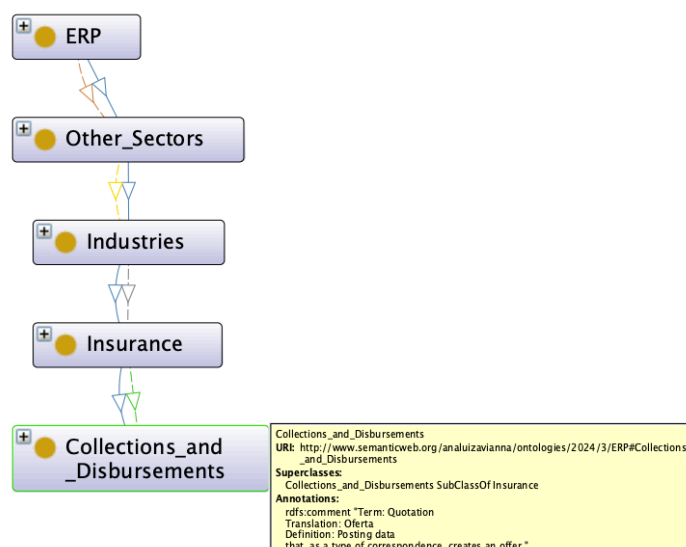
Fonte: Elaborado pela autora.

Figura 82 - *Subframe* Marketing\_Accounting\_and\_Pricing

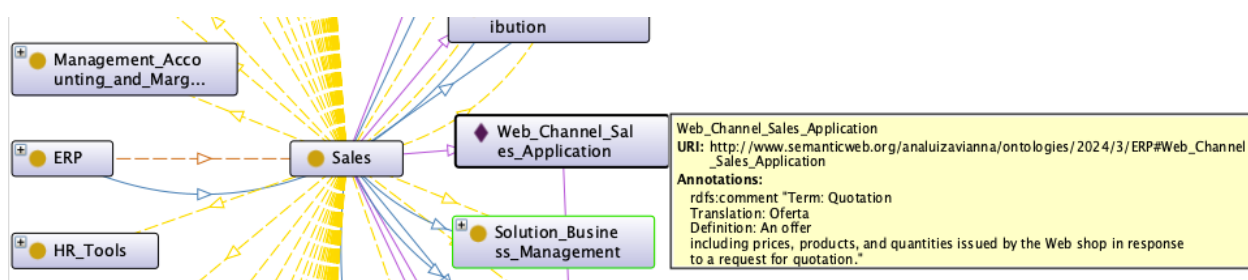


Fonte: Elaborado pela autora.

Com relação ao termo ‘oferta’, encontramos duas definições na base terminológica. A primeira é definida como uma oferta que inclui preços, produtos e quantidades emitidas pela loja em resposta a um pedido de orçamento. A segunda se refere à publicação de dados que, como forma de correspondência, geram uma oferta. Tal unidade está localizada no componente *Other Sectors*, mais especificamente no *subframe* *Collections\_and\_Disbursements*, na área de *insurance*, e no componente *Sales*, na instância *Web\_Channel\_Sales\_Application*, como ilustram as Figuras 83 e 84.

Figura 83 - *Subframe* Collections\_and\_Disbursements

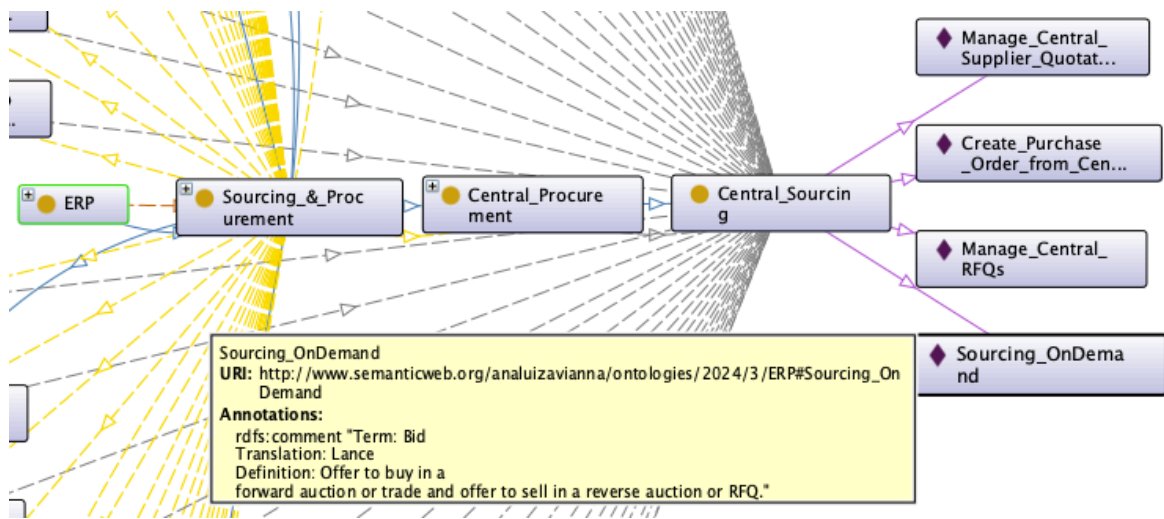
Fonte: Elaborado pela autora.

Figura 84 - Anotação instância *Web\_Channel\_Sales\_Application*

Fonte: Elaborado pela autora.

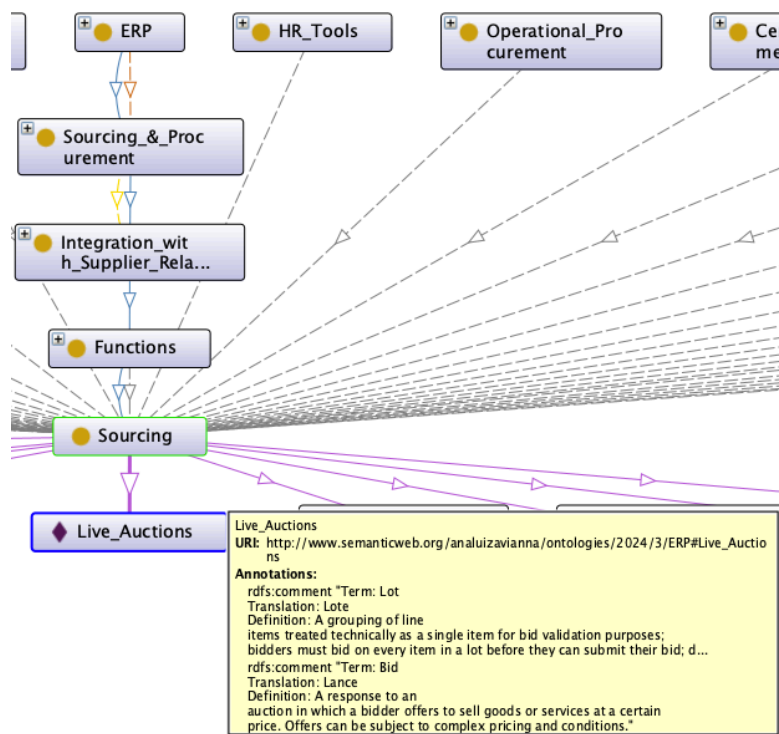
O último termo, ‘lance’, apresenta três definições na base terminológica: (i) oferece-se para comprar em um leilão ou negociação a termo, bem como se oferece para vender em um leilão reverso, ou em solicitação de cotação (*request for quotation*); (ii) uma oferta para comprar a um preço específico em um leilão ou negociação a termo, ou uma oferta para vender a um preço específico em um leilão reverso ou em solicitação de cotação; (iii) uma resposta a um leilão em que um licitante se oferece para vender bens ou serviços a um determinado preço. As ofertas podem estar sujeitas a preços e condições. Essa unidade lexical se encontra no componente *Sourcing & Procurement*, no *subframe* *Sourcing*, instância *Live\_Auctions*, e no *subframe* *Central\_Sourcing*, instância *Sourcing\_OnDemand*, e no componente *Sales*, *subframe* *Sales\_Configuration*, instância *Dynamic\_Pricing\_Engine*, como vemos por meio das Figuras 85, 86, 87.

Figura 85 - Anotação instância *Sourcing\_OnDemand*

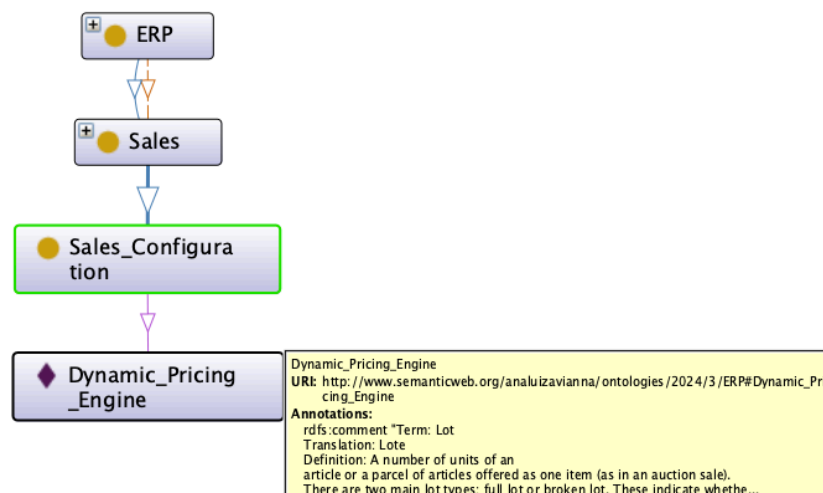


Fonte: Elaborado pela autora.

Figura 86 - Anotação instância *Live\_Auctions*



Fonte: Elaborado pela autora.

Figura 87 - Anotação instância *Dynamic\_Pricing\_Engine*

Fonte: Elaborado pela autora.

Para a harmonização desse conjunto de termos, selecionamos ‘cotação’ como termo principal, visto que ele é frequente em diversos componentes e *subframes* do *Knowledge Graph*. O termo ‘cotação’ é mais abrangente, sobretudo por ser utilizado em situação empresarial para busca de fornecedores e análise de preços e produtos. Ainda que ‘oferta’ seja usado como referência para tratar da cotação, percebe-se que o termo ‘oferta’ não pode ser utilizado em contexto de solicitação de cotação, por exemplo, visto que seu conceito não está associado, especificamente, à solicitação RFQ como cotação. Outro ponto se refere à relação entre os dois termos. A cotação depende da oferta e demanda. Quando se tem uma oferta, exige-se uma cotação de preços para definir a compra de produtos de um determinado fornecedor. Caso a oferta esteja alta, a cotação cai; se a procura for alta, o preço aumenta. Embora tenhamos essa relação entre a terminologia, podemos observar pelas definições que, na conjuntura em que as unidades estão inseridas, ‘oferta’ se aproxima mais do conceito de cotação e, por isso, ‘cotação’ se torna o termo principal.

Com relação às variantes, como dito anteriormente, o termo ‘oferta’ é uma variação de cotação por não ser um termo tão amplo a ponto de comportar o contexto de solicitação de cotação e estar mais próximo apenas da situação de venda. O termo ‘lance’, por sua vez, não foi escolhido como termo principal devido a sua especificidade nas definições. A unidade ‘lance’ está muito associada ao contexto de leilão, embora também apresente uma proximidade com a solicitação de cotação. Do nosso ponto de vista, ‘lance’ é uma etapa posterior à cotação, posto que, para que haja um leilão (ou pregão), é necessário, primeiro, haver a cotação para que o lance seja dado.



Considerando a harmonização realizada para esse conjunto de termos – cotação, lance e oferta –, elaboramos a seguinte ficha terminológica (Quadro 26).

Quadro 26 - Ficha terminológica do termo ‘cotação’

<b>Termo:</b> Cotação
<b>Entrada:</b> Cotação
<b>Classe gramatical:</b> sf.
<b>Abreviatura/sigla:</b> Cot.
<b>Domínio:</b> Supply Chain
<b>Definição:</b> Cotação é uma oferta efetuada por um fornecedor a uma instituição para o fornecimento de materiais e para a prestação de serviços sob determinados termos e condições, como preços, datas de entregas e pagamentos. É, também, uma resposta de um determinado fornecedor a uma RFQ ( <i>request for quotation</i> – solicitação de cotação).
<b>Fonte da definição:</b> Elaborada a partir das definições da base terminológica.
<b>Contexto:</b> Supply Chain, Sourcing & Procurement e Sales.
<b>Variantes:</b> Oferta, Lance
<b>Equivalentes:</b> Quotation, Bid

Fonte: Elaborado pela autora.

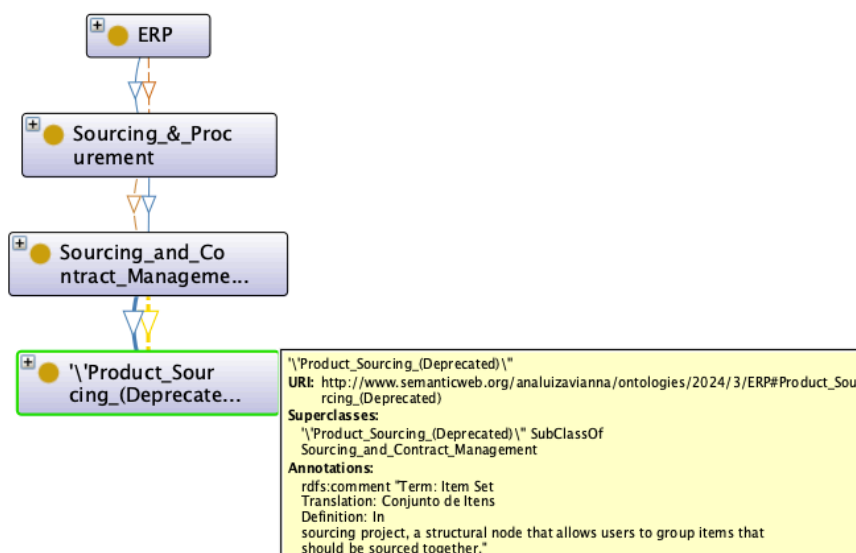
A ficha terminológica do conjunto de termos composto por ‘cotação’, ‘lance’ e ‘oferta’ apresenta como termo principal ‘cotação’. Além disso, também se encontra a classe gramatical, a abreviatura da unidade, que poderia ser adotada pela empresa para ser utilizada em documentos e bases de dados, o domínio ao qual o termo pertence, a definição, a fonte da definição, as variantes e, por fim, os equivalentes em língua inglesa.

### 6.2.3 Conjunto de termos 3: Conjunto de itens

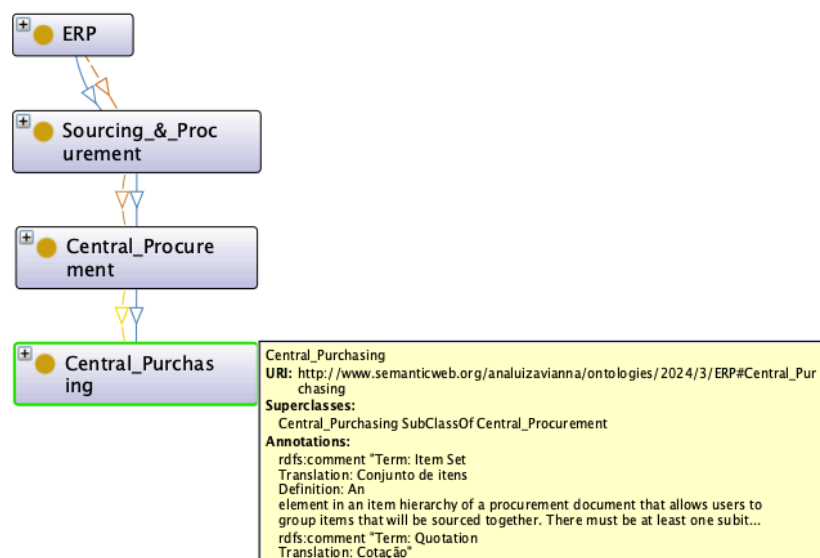
Nosso terceiro grupo de termos para análise é composto pelas seguintes unidades lexicais: conjunto de itens, pacote e lote. Tais termos estão inseridos em diferentes áreas do ERP e são utilizados de acordo com seu contexto específico. Para a harmonização desse conjunto, buscamos as definições na base terminológica e em dicionários especializados e de língua geral.

O termo ‘conjunto de itens’ apresentou duas definições na base terminológica. A primeira está relacionada a um projeto de fornecimento (*Sourcing Project*) no qual os usuários podem reunir os itens que necessitam ser adquiridos. A segunda definição se refere a um elemento de uma hierarquia de itens que será adquirido a partir de um documento de aquisição. Nesse sentido, deverá haver, pelo menos, um subitem no conjunto de itens sem apresentar processos subsequentes, visto que é utilizado apenas para agrupar subitens. Vale destacar que ‘conjunto de itens’ não apresentou resultados em nossas buscas dicionarísticas, uma vez que tal unidade não é, até certo ponto, tão especializada. Nossa busca pelas palavras ‘conjunto’ e ‘item’ no dicionário Caldas Aulete não resultou em definições próximas das apresentadas na base terminológica. No que tange à localização de ‘conjunto de itens’ no *Knowledge Graph* e nos *frames*, uma das ocorrências está situada no componente *Sourcing & Procurement*, no *subframe* *Product\_Sourcing* (Figura 88), e a segunda ocorrência também está no componente *Sourcing & Procurement*, no *subframe* *Central\_Purchasing* (Figura 89).

Figura 88 - *Subframe* *Product\_Sourcing*



Fonte: Elaborado pela autora.

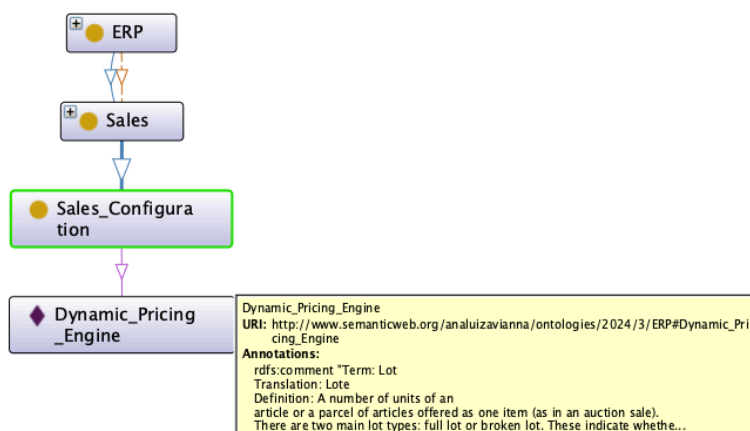
Figura 89 - *Subframe* Central\_Purchasing

Fonte: Elaborado pela autora.

No que concerne ao termo ‘lote’, encontramos sete definições na base terminológica. A primeira se refere à quantidade de um objeto como unidade de estoque, produção, venda, amostragem ou transporte. Os itens de um lote são de uma única classe, composição, modelo, tamanho e tipo, ou seja, são produzidos seguindo as mesmas condições e características dentro de limites especificados. A segunda definição condiz com o número de unidades de um artigo ou de um pacote de objetos oferecidos como um único item de, por exemplo, em uma venda em leilão. Encontramos dois tipos de lote, lote inteiro ou lote quebrado, que indicam como os itens podem ou não ser leiloados, se em sua totalidade, ou em parte. A terceira definição diz respeito a uma quantidade de objetos recebida em estoque ou produzida, apresentando os mesmos custos e especificações de produção. Para a quarta definição, ‘lote’ se refere ao agrupamento de todos os materiais de uma entrega para que as mercadorias possam ser rastreadas e debitadas/creditadas em lojas. Como quinta definição, o termo ‘lote’ se trata de uma quantidade de formulários pré-numerados vendidos em lojas de documentos. A sexta definição remete a um conjunto de materiais semelhantes, um único material ou um pacote de materiais que quantificam como um e que apresentam um número de identificação. Por fim, lote também pode ser definido como um conjunto de itens de linha identificados como um único item para validação de proposta. Nesse caso, o licitante deve dar lances em todos os itens de um lote antes de enviar sua proposta. Assim, os dados são acumulados até o nível do lote para validar a oferta. Considerando as diferentes definições de ‘lote’, buscamos no Dicionário do Marcílio (2024)

descrições sobre a unidade. A primeira está relacionada ao lote de produtos e lote de ações, a segunda acepção se refere a um lote de produtos (*batch*) e ao mercado de capitais (*lot*). Na busca pela definição de ‘lote’ em um dicionário de língua geral, encontramos a seguinte definição: “grupo de objetos, mercadorias etc. do mesmo tipo ou incluídos na mesma operação (lote de medicamentos).” (Caldas Aulete, 2009)<sup>62</sup>. Com relação à localização de ‘lote’ no KG, as ocorrências situam-se nos componentes *Sales*, *subframe Sales\_Distribution* e instância *Dynamic\_Pricing\_Engine*, *Manufacturing*, *subframe Material\_Requirements\_Planning*, *Professional Services*, *subframe Basic\_Functions* e instância *Manufacturing\_Execution*, *Sourcing & Procurement*, *subframe Sourcing* e instância *Live\_Auctions*, *Supply Chain*, *subframe Logistics\_Cross\_Topics* e instância *Retail*, *Supply Chain*, *subframe Warehousing* e instância *Basic\_Integration\_with\_Extended\_Warehouse\_Management*, conforme podemos observar nas figuras 90, 91, 92, 93, 94, 95 e 96.

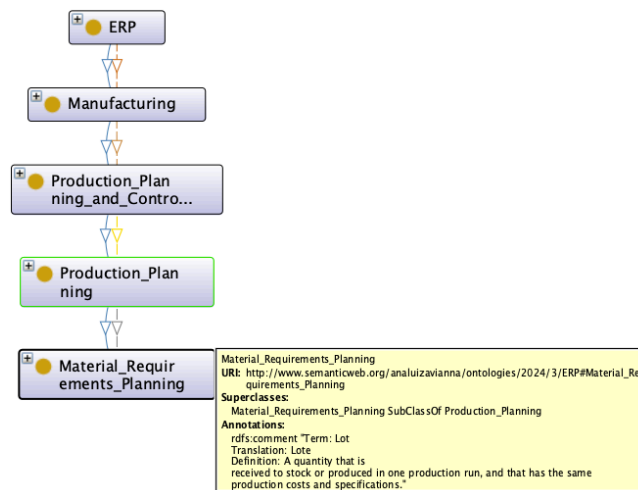
Figura 90 - Anotação instância *Dynamic\_Pricing\_Engine*



Fonte: Elaborada pela autora.

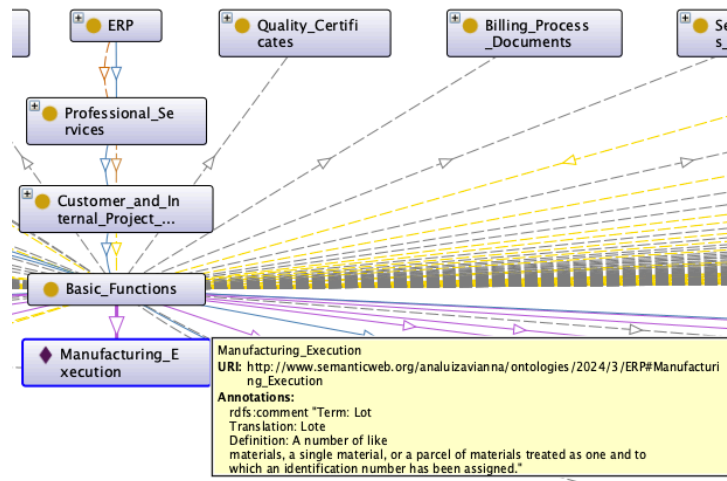
<sup>62</sup> LOTE. *In.*: Dicionário Caldas Aulete Online. Disponível em: <https://aulete.com.br/lote> Acesso em: 1 maio 2024.

Figura 91 - *Subframe* Material\_Requirements\_Planning



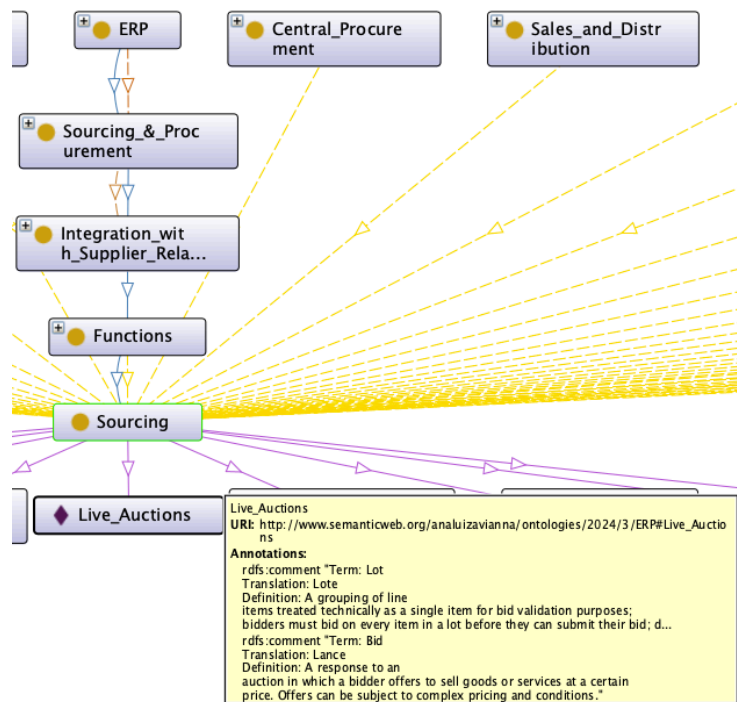
Fonte: Elaborada pela autora.

Figura 92 - Anotação instância *Manufacturing\_Execution*



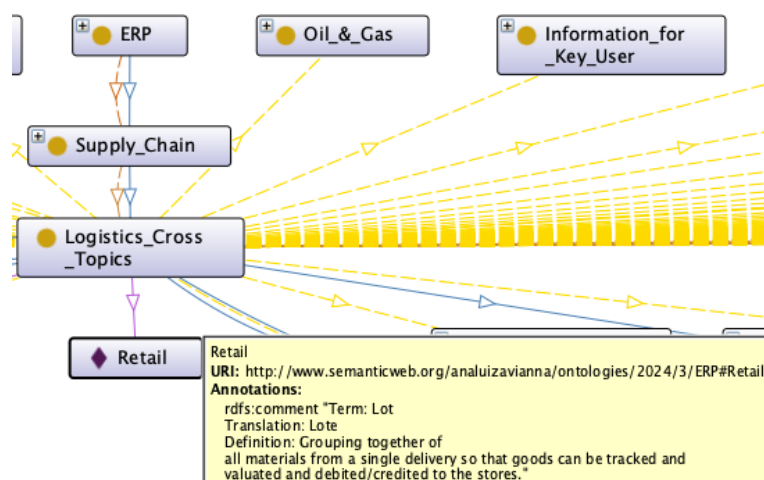
Fonte: Elaborada pela autora.

Figura 93 - Anotação instância *Live\_Auctions*

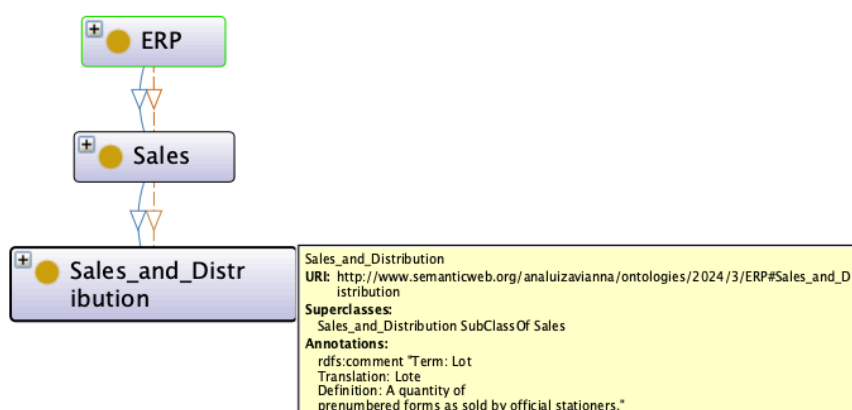


Fonte: Elaborada pela autora.

Figura 94 - Anotação instância *Retail*

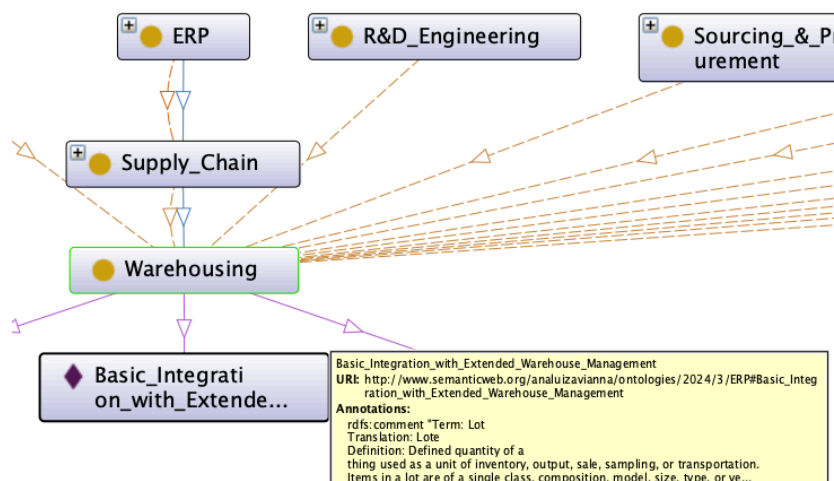


Fonte: Elaborada pela autora.

Figura 95 - *Subframe Sales\_and\_Distribution*

Fonte: Elaborado pela autora.

Figura 96 - Anotação instância  
*Basic\_Integration\_with\_Extended\_Warehouse\_Management*



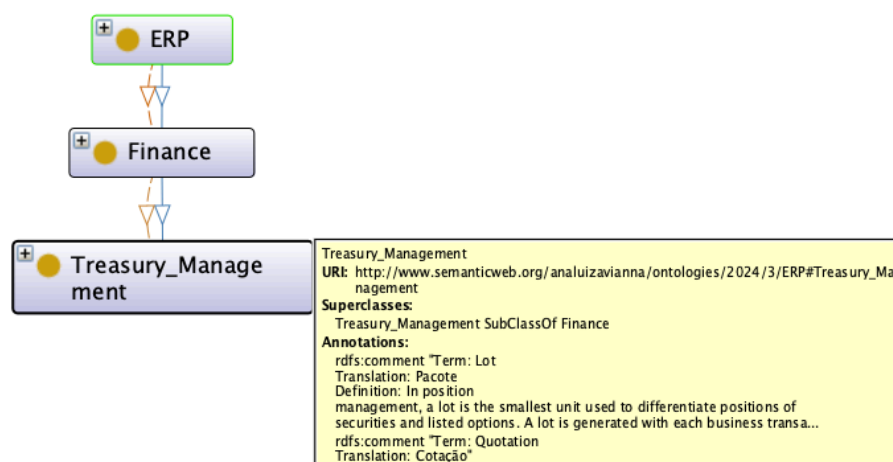
Fonte: Elaborada pela autora.

Nas Figuras 91, 92, 93, 94, 95, 96 e 97, podemos observar todas as ocorrências do termo ‘lote’ no ERP. Desse modo, ‘lote’ está localizado em diferentes componentes do *software*, como em *Sales*, *subframe Sales\_Distribution* e instância *Dynamic\_Pricing\_Engine*, *Manufacturing*, *subframe Material\_Requirements\_Planning*, *Professional\_Services*, *subframe Basic\_Functions* e instância *Manufacturing\_Execution*, *Sourcing & Procurement*, *subframe Sourcing* e instância *Live\_Auctions*, *Supply Chain*, *subframe Logistics\_Cross\_Topics* e instância *Retail*, *Supply Chain*, *subframe Warehousing* e instância *Basic\_Integration\_with\_Extended\_Warehouse\_Management*.

A última unidade desse conjunto de termos é ‘pacote’, para a qual encontramos a seguinte definição: o pacote é a menor unidade utilizada para diferenciar posições de

títulos e opções listadas, sendo gerada a cada transação comercial que aumenta uma posição. No que diz respeito à definição de ‘pacote’ no dicionário Caldas Aulete (2009)<sup>63</sup>, observamos que a palavra se refere a qualquer embrulho ou conjunto de objetos não muito grandes reunidos, formando um só volume, assim como um conjunto de elementos contidos em um pacote. Com relação à localização do termo no *Knowledge Graph*, a unidade se situa no componente de *Finance*, *subframe* *Treasury\_Management* (Figura 97).

Figura 97 - *Subframe* *Treasury\_Management*



Fonte: Elaborada pela autora.

Considerando as definições anteriores e o contexto do KG e dos *frames*, selecionamos o termo ‘conjunto de itens’ como termo principal, e ‘lote’ e ‘pacote’ como variantes. Ao analisarmos as definições de ‘conjunto de itens’, percebemos que ele está relacionado com diferentes produtos/objetos/itens que uma empresa produz ou vende. Ainda que ele esteja presente apenas no componente *Sourcing & Procurement*, pode-se entender que tal termo é mais amplo. No caso de ‘lote’, as definições também se referem a um conjunto de itens produzidos por uma empresa atentando para a linha de produção e características uniformes. Embora seja um termo recorrente em muitos componentes do ERP, ao analisarmos o contexto de cada ocorrência, observamos que o termo ‘lote’ é mais utilizado em domínios específicos como na indústria farmacêutica, no varejo e no planejamento de materiais. Devido a isso, ‘lote’ não poderia ser considerado o termo principal, tendo em vista as especificidades que as definições apresentam. No caso do termo ‘pacote’, a única definição mais especializada se refere

<sup>63</sup> PACOTE. In.: Dicionário Caldas Aulete Online. Disponível em: <https://aulete.com.br/pacote> Acesso em: 1 maio 2024.



ao domínio financeiro, fato que nos impede de selecionar a unidade como termo principal. Ao examinarmos todas as definições e os conceitos dos termos, entendemos que ‘conjunto de itens’ pode ser considerado como um termo guarda-chuva, uma vez que é uma unidade mais ampla e que pode abarcar todas as outras definições mais específicas.

A partir da seleção do termo principal e das variantes para o conjunto composto por ‘conjunto de itens’, ‘lote’ e ‘pacote’, elaboramos a seguinte ficha terminológica, como mostra o Quadro 27.

Quadro 27 - Ficha terminológica do termo ‘conjunto de itens’

<b>Termo:</b> Conjunto de itens
<b>Entrada:</b> Conjunto de itens
<b>Classe gramatical:</b> Loc.
<b>Abreviatura/sigla:</b> Conj. de itens
<b>Domínio:</b> Sourcing & Procurement
<b>Definição:</b> Grupo de produtos, os quais apresentam as mesmas características e condições, que são adquiridos a partir de um documento de aquisição, seguindo uma hierarquia de itens e subitens.
<b>Fonte da definição:</b> Elaborado a partir das definições da base terminológica.
<b>Contexto:</b> Sourcing & Procurement
<b>Variantes:</b> Lote e Pacote
<b>Equivalentes:</b> Item set e Lot

Fonte: Elaborado pela autora.

A ficha terminológica do conjunto de termos composto por ‘conjunto de itens’, ‘lote’ e ‘pacote’ apresenta como termo principal ‘conjunto de itens’. O termo é então seguido da sua classe gramatical, a abreviatura que a empresa pode adotar para ser utilizada posteriormente em documentos e bases de dados, o domínio ao qual a unidade pertence, uma definição, a fonte da definição, o contexto de uso, as variantes e, finalmente, os equivalentes em inglês.

Após a aplicação da nossa metodologia semântico-terminológica para harmonização de termos e conceitos, percebemos que a camada contextual fornecida pelos *frames* e pelo *Knowledge Graph* contribuiu para a escolha do termo principal e

das variantes. Inicialmente, tinha-se em mente que a escolha poderia ser feita pela frequência e pela quantidade de definições que determinado termo apresenta. Entretanto, com as informações extras advindas da representação do conhecimento, percebe-se que a localização do termo no mapa/KG permite uma compreensão melhor do domínio e, conseqüentemente, do comportamento da terminologia, facilitando o processo de tomada de decisão.

Um ponto interessante a se destacar diz respeito às informações provenientes não apenas do *corpus*, mas do material de apoio – dicionários, glossários etc. – que auxiliaram no entendimento dos componentes do ERP e dos termos *per se*. No caso do termo ‘empregado’, para selecioná-lo como termo principal, consideramos as informações dos *frames* e do KG, assim como das definições e da legislação brasileira. Isso ocorreu devido à necessidade de se ter mais dados para realmente estabelecer as relações entre termo e conceito e as associações do termo com as demais unidades. À medida em que apenas as definições da base terminológica não fornecem uma contextualização mais aprofundada (ou, em alguns casos, não tão abrangente), recorrer a outros materiais se torna imprescindível para a tarefa de harmonização.

Durante esse processo de aplicação e análise, podemos perceber que investigar como funciona o processo de cada componente do ERP contribuiu para a escolha dos termos. No conjunto de termos 2 – ‘cotação’, ‘oferta’ e ‘lance’ –, observamos que o entendimento acerca do conceito de demanda e oferta, de leilão e dos documentos de cotação foram fundamentais para que se estabelecessem o termo principal e as variantes, sobretudo quando analisamos o termo ‘lance’.

Por fim, vale salientar a complexidade em se harmonizar termos que não apresentam características tão distintas entre eles, como, por exemplo, ‘empregado’ e ‘funcionário’, ‘conjunto de itens’ e ‘lote’ e ‘oferta’ e cotação’. Para a escolha dos termos principais, nessas situações, considerou-se não só a recorrência no *software* (vide que ‘conjunto de itens’ não é tão frequente quanto ‘lote’), mas a abrangência do termo. Não poderíamos selecionar um termo muito específico (ou de um *subframe* ou instância tão exclusivo), visto que o termo principal deve, a princípio, abranger todos os termos sinônimos a ele. ‘Conjunto de itens’, por exemplo, foi considerado como um termo guarda-chuva, pois, ainda que não seja tão frequente, é capaz de abarcar os demais termos sem ser de natureza tão singular.

Na próxima seção, apresentaremos os resultados da aplicação da tarefa de harmonização feita com o ChatGPT, considerando as estratégias de *prompt* discutidas no capítulo metodológico.

### 6.3 LLMs e Harmonização Terminológica

Após aplicarmos a nossa metodologia semântico-terminológica para harmonização de termos e conceitos de forma manual, isto é, termos seguido os passos metodológicos para examinarmos a funcionalidade da abordagem, nesta seção, analisaremos os resultados gerados pelo ChatGPT acerca da tarefa de harmonização. Inicialmente, apresentaremos as perguntas que criamos para utilizarmos no *prompt* para a tarefa. Em seguida, exploraremos as respostas geradas, relacionando com as estratégias usadas e aproximando a harmonização feita pela máquina com a nossa harmonização. Para essa análise, selecionamos o conjunto de termos 1, composto por Empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal.

#### 6.3.1 Estratégias de *prompt* para a tarefa de harmonização de termos e conceitos

Com a transformação digital e a revolução na área de Inteligência Artificial, sobretudo com o crescente uso de modelos generativos, busca-se desenvolver e aperfeiçoar estratégias para que *Large Language Models* (LLMs) possam desempenhar tarefas cognitivas humanas. Para que seja possível realizar alguma atividade, a interação entre modelo de linguagem e ser humano é feita por meio de *prompts*. Assim, o indivíduo fornece uma instrução a fim de que o modelo gere uma resposta adequada ao questionamento. No capítulo metodológico, tratamos dos *prompts* e como eles são pensados para obter os melhores resultados do modelo de linguagem. Uma área que se aprofunda no desenvolvimento de *prompts* é a Engenharia de *Prompt*, que busca planejar, processar e otimizar *prompts* para que o LLM seja instruído e que forneça a resposta mais próxima, adequada e precisa possível. Para que isso ocorra, é preciso delimitar o objetivo da tarefa para que as instruções sejam criadas para atingir a meta.

No contexto desta tese, com relação às perguntas a serem feitas no *prompt*, como estratégia, iniciamos com instruções mais amplas, questionando o que é a harmonização terminológica, a fim de compreendermos qual abordagem o modelo segue para essa

tarefa. Em seguida, nosso segundo *prompt* focou em pedir a harmonização do conjunto de termos, sem considerar exemplos e definições. Para a terceira pergunta, adicionamos definições sem exemplos ao conjunto de termos e, no quarto *prompt*, acrescentamos exemplos. O quinto *prompt* desenvolvido para a tarefa de harmonização objetivou trabalhar com o *Knowledge Graph* no LLM para que todo o contexto da representação do conhecimento e dos *frames* fosse considerado na execução da tarefa. Para isso, verbalizamos o KG para que fosse possível adicionarmos essa camada de informação advinda do KG do ERP. Por fim, desenvolvemos um sexto *prompt*, unindo as definições da base terminológica com exemplos e a representação do conhecimento com o objetivo de fornecer mais camadas de informação para o modelo de linguagem.

A fim de analisarmos os resultados gerados pelo ChatGPT, selecionamos o conjunto de termos composto por ‘empregado’, ‘funcionário’, ‘colaborador’, ‘trabalhador temporário’, ‘recurso externo’, ‘usuário de colaboração’, ‘operário’, ‘trabalhador’ e ‘membro do pessoal’. Considerando os termos a serem investigados, criamos os seguintes *prompts*<sup>64</sup>:

Quadro 28 - *Prompts* para a tarefa de harmonização de termos e conceitos.

**Prompt 1:**

O que é harmonização de termos e conceitos?

**Prompt 2:**

Considerando a harmonização de termos e conceitos, como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal. Qual seria o termo principal e quais seriam as variantes?

**Prompt 3:**

Considere as seguintes definições:

Uma pessoa que contribui ou contribuiu para a criação de bens e serviços na empresa com base em um contrato de trabalho, um acordo sem aviso prévio ou um contrato de serviços.

Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços para uma empresa. O funcionário inclui funcionários internos e externos (executores de serviços).

Diferentemente dos externos, o funcionário interno está vinculado a instruções e é obrigado a aderir às políticas e aos regulamentos da empresa. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços na empresa com base em um contrato de trabalho para serviços.

Um funcionário que não tem um contrato explícito de emprego permanente ou de longo prazo. Um indivíduo que realiza trabalhos para uma organização em caráter não permanente.

Um recurso externo que é um funcionário terceirizado contratado como funcionário externo.

<sup>64</sup> A elaboração dos *prompts* para esta análise de dados foi abordada na seção 5.3.1 Terminologia e LLMs: Instruções e estratégias de *prompt* para a tarefa de harmonização terminológica.

Um usuário de colaboração que é um funcionário terceirizado que participa de um cenário de colaboração.

Um trabalhador que pode operar um dispositivo ou trabalhar sem um dispositivo.

Funcionário ou trabalhador externo.

Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando as definições acima? Qual seria o termo principal e quais seriam as variantes?

**Prompt 4:**

Considere as seguintes definições:

Empregado é uma pessoa que tem um contrato de trabalho, um acordo sem aviso prévio ou um contrato de serviços. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços para uma empresa.

O funcionário inclui funcionários internos e externos (executores de serviços).

Diferentemente dos externos, o funcionário interno está vinculado a instruções e é obrigado a aderir às políticas e aos regulamentos da empresa. Uma pessoa que contribui ou contribuiu para a criação de bens ou serviços na empresa com base em um contrato de trabalho para serviços. O funcionário constitui o principal foco de interesse da maioria dos componentes de aplicação de recursos humanos, como administração de pessoal, gerenciamento de tempo, folha de pagamento e assim por diante. Todos os processos de negócios nesses componentes de aplicação estão relacionados ao funcionário.

Um funcionário que não tem um contrato explícito de emprego permanente ou de longo prazo.

Um indivíduo que trabalha para uma organização de forma não permanente. Os trabalhadores temporários são geralmente conhecidos como profissionais independentes, trabalhadores com contrato temporário, empreiteiros ou consultores.

Um recurso externo que é um funcionário terceirizado contratado como funcionário externo.

Por exemplo, consultores ou auditores externos. Portanto, eles compartilham o endereço de trabalho do empregador contratado por um determinado período de tempo.

Um usuário de colaboração que é um funcionário terceirizado que participa de um cenário de colaboração, por exemplo, inventário gerenciado pelo fornecedor.

Um usuário de colaboração precisa de acesso a um sistema, por exemplo, para manter estoques.

Um trabalhador pode operar um dispositivo ou trabalhar sem um dispositivo. Exemplo: motorista de empilhadeira ou selecionador. Um operador no chão de fábrica que monta placas, unidades ou outros materiais que você fabrica.

Funcionário ou trabalhador externo.

Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando as definições acima? Qual seria o termo principal e quais seriam as variantes?

Fonte: Elaborado pela autora.

Com relação aos *prompts* 5 e 6, é importante destacarmos alguns pontos. Dentre as distintas técnicas de *prompt*, fornecer a contextualização das informações de um *Knowledge Graph* para o LLM no pós-processamento é uma forma de propiciar dados externos para o modelo com o objetivo de que ele considere as informações extras juntamente com os seus dados de treinamento para gerar uma resposta mais precisa e adequada (cf. Baldazzi *et al.*, 2023; Hertling; Paulheim, 2023; Bontcheva; Davis, 2009).

Em nosso trabalho, após a elaboração do KG do ERP, optamos por verbalizar a arquitetura para adicionarmos a camada de informação proveniente do *Knowledge Graph* no LLM. A verbalização é uma forma de tornar as informações ontológicas linguisticamente mais aceitáveis (Kaljurand; Fuchs, 2007; Mellish; Sun, 2005).

Para essa tarefa, geralmente, buscam-se subsídios no *Attempto Controlled English (ACE) Language Manual* (Fuchs; Schwertel; Schwitter, 1999), que descreve um conjunto de princípios de construção e interpretação para a escrita de especificações. O ACE tem como objetivo principal elaborar estruturas linguísticas adequadas para a língua inglesa a fim de reduzir a ambiguidade e vagueza, tornando as estruturas processáveis por máquina e executáveis. Para a nossa verbalização, tomando por base as informações do *Knowledge Graph* e sua robustez, verbalizamos apenas as partes em que os termos – ‘empregado’, ‘colaborador’, ‘funcionário’, ‘trabalhador’, ‘trabalhador temporário’, ‘operário’, ‘recurso externo’, ‘usuário de colaboração’ e ‘membro do pessoal’ – ocorriam no grafo.

Em segundo lugar, vale destacar que a nossa verbalização apresenta informações terminológicas em inglês e português, uma vez que nosso *Knowledge Graph* foi feito em língua inglesa (como já justificado anteriormente, nossa intenção é tornar o KG disponível e, por isso, optamos por desenvolvê-lo em inglês). Nesse sentido, descrevemos as informações, mantendo os termos em inglês e colocando as devidas traduções entre parênteses.

O último ponto a ser evidenciado condiz com o modelo em que o *prompt 6* foi aplicado. Para essa instrução, tendo em vista a quantidade de informações explicitadas, o modelo GPT-4 32k não conseguiu processar todo o conteúdo. Logo, apenas para o *prompt 6*, utilizamos o modelo GPT-4 para avaliarmos a eficácia da instrução e dos resultados. Para essa etapa, também testamos o modelo nas temperaturas 0 e 1. Os detalhes do *Prompt 5* e *6* são apresentados na sequência.

#### Quadro 29 - *Prompts* desenvolvidos para o modelo GPT-4

**Prompt 5:**

Considere a seguinte verbalização:

ERP é um software de gerenciamento de empresas.

ERP tem cross components.

Cross components tem o termo Employee (empregado).

ERP tem componentes.

Human Resources é um componente do ERP.

Human Resources tem recursos.

Payroll, Personnel Management e HR Master Data são recursos de Human Resources.

Payroll tem o termo Worker (operário).  
 HR Master Data tem as instâncias HCM\_Employee\_Central e Business\_User.  
 A instância HCM\_Employee\_Central tem os termos Contingent Worker (trabalhador temporário) e Employee (colaborador).  
 A instância Business User tem os termos Collaboration User (usuário de colaboração), External Resource (recurso externo) e Worker (trabalhador).  
 Personnel Management tem termos.  
 Employee (empregado) é um termo de Personnel Management.  
 Personnel Management tem funcionalidades.  
 Budget Management e Personnel Administration são funcionalidades de Personnel Management.  
 Budget Management tem o termo Employee (empregado).  
 Personnel Administration tem os termos Contingent Worker (trabalhador temporário) e External Resource (recurso externo).  
 ERP inclui Other Sectors.  
 Other Sectors tem Industries.  
 Industries tem Public Sector.  
 Public Sector tem recursos.  
 Public Sector Management é um recurso de Public Sector.  
 Investigative Case Management é uma funcionalidade de Public Sector Management.  
 Investigative Case Management tem o termo Employee (empregado).  
 Supply Chain é um componente do ERP.  
 Supply Chain tem recursos.  
 Logistic Cross Topics é um recurso de Supply Chain.  
 Logistic Cross Topics tem funcionalidades.  
 Logistic Execution é uma funcionalidade de Logistic Cross Topics.  
 Logistic Execution tem app para gerenciamento de Task and Resource Management  
 Task and Resource Management tem o termo Worker (funcionário).  
 Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando a verbalização acima? Qual seria o termo principal e quais seriam as variantes?

**Prompt 6:**

Considere as seguintes definições:  
 Empregado é uma pessoa que contribui ou contribuiu para a criação de bens e serviços na empresa com base em um contrato de trabalho, um acordo sem aviso prévio ou um contrato de serviços.  
 O funcionário inclui funcionários internos e externos (executores de serviços). Diferentemente dos externos, o funcionário interno está vinculado a instruções e é obrigado a aderir às políticas e aos regulamentos da empresa. O funcionário constitui o principal foco de interesse da maioria dos componentes de aplicação de recursos humanos, como administração de pessoal, gerenciamento de tempo, folha de pagamento e assim por diante.  
 Os trabalhadores temporários são geralmente conhecidos como profissionais independentes, trabalhadores com contrato temporário, empreiteiros ou consultores.  
 Um recurso externo é um funcionário terceirizado contratado como funcionário externo. Por exemplo, consultores ou auditores externos. Portanto, eles compartilham o endereço de trabalho do empregador contratado por um determinado período de tempo.  
 Um usuário colaboração que é um funcionário terceirizado que participa de um cenário de colaboração, por exemplo, inventário gerenciado pelo fornecedor.  
 Um operário pode operar um dispositivo ou trabalhar sem um dispositivo. Exemplo: motorista de empilhadeira ou selecionador. Um operador no chão de fábrica que monta placas, unidades ou outros materiais que você fabrica.  
 Considere a seguinte verbalização:

ERP é um software de gerenciamento de empresas.  
 ERP tem cross components.  
 Cross components tem o termo Employee (empregado).  
 ERP tem componentes.  
 Human Resources é um componente do ERP.  
 Human Resources tem recursos.  
 Payroll, Personnel Management e HR Master Data são recursos de Human Resources.  
 Payroll tem o termo Worker (operário).  
 HR Master Data tem as instâncias HCM\_Employee\_Central e Business\_User.  
 A instância HCM\_Employee\_Central tem os termos Contingent Worker (trabalhador temporário) e Employee (colaborador).  
 A instância Business User tem os termos Collaboration User (usuário de colaboração), External Resource (recurso externo) e Worker (trabalhador).  
 Personnel Management tem termos.  
 Employee (empregado) é um termo de Personnel Management.  
 Personnel Management tem funcionalidades.  
 Budget Management e Personnel Administration são funcionalidades de Personnel Management.  
 Budget Management tem o termo Employee (empregado).  
 Personnel Administration tem os termos Contingent Worker (trabalhador temporário) e External Resource (recurso externo).  
 ERP inclui Other Sectors.  
 Other Sectors tem Industries.  
 Industries tem Public Sector.  
 Public Sector tem recursos.  
 Public Sector Management é um recurso de Public Sector.  
 Investigative Case Management é uma funcionalidade de Public Sector Management.  
 Investigative Case Management tem o termo Employee (empregado).  
 Supply Chain é um componente do ERP.  
 Supply Chain tem recursos.  
 Logistic Cross Topics é um recurso de Supply Chain.  
 Logistic Cross Topics tem funcionalidades.  
 Logistic Execution é uma funcionalidade de Logistic Cross Topics.  
 Logistic Execution tem app para gerenciamento de Task and Resource Management  
 Task and Resource Management tem o termo Worker (funcionário).  
 Como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e membro do pessoal, considerando as definições e a verbalização acima. Qual seria o termo principal e quais seriam as variantes?

Fonte: Elaborado pela autora.

Após descrevermos os seis *prompts* elaborados para a realização dos experimentos com os modelos de linguagem GPT-4 32k e, para o *prompt* 6, GPT-4, a seguir, apresentaremos os resultados obtidos para cada *prompt*.

### 6.3.2 Conjunto de termos: Empregado

Sendo um dos nossos objetivos treinar um LLM (*Large Language Model* – grande modelo de linguagem) para desempenhar a tarefa de harmonização



automaticamente e, para fins de aplicação da nossa metodologia semântico-terminológica para harmonização de termos e conceitos, desenvolvemos *prompts* para avaliarmos o modelo e alcançarmos resultados mais precisos. Como vimos anteriormente, elaboramos seis *prompts* para investigarmos como a harmonização é feita e, a cada instrução, inserimos mais uma camada de informação até implementarmos o *Knowledge Graph* arquitetado para o nosso propósito metodológico. Para esta análise, avaliaremos os resultados gerados pelos modelos GPT-4 32k e GPT-4 em duas temperaturas, 1 e 0.

Com relação à primeira pergunta, ‘O que é harmonização de termos e conceitos?’, o modelo apresentou duas respostas distintas devido à temperatura selecionada, como podemos ver nos Quadros 30 e 31.

Quadro 30 - Resposta *prompt* 1 - GPT-4 32k temperatura<sup>65</sup> 0

A harmonização de termos e conceitos é um processo que busca garantir que diferentes pessoas, grupos ou sistemas usem e entendam termos e conceitos da mesma maneira. Isso é especialmente importante em áreas como a ciência da informação, a gestão de dados, a pesquisa e o desenvolvimento, onde a comunicação clara e precisa é essencial.

Por exemplo, em uma organização grande, diferentes departamentos podem usar termos diferentes para se referir à mesma coisa. Um departamento pode se referir a "clientes" enquanto outro usa o termo "contas". Isso pode levar a confusão e ineficiência. A harmonização de termos e conceitos envolveria a criação de um vocabulário comum que todos os departamentos usariam.

Da mesma forma, na pesquisa científica, diferentes pesquisadores podem usar termos e conceitos de maneira ligeiramente diferente. A harmonização nesse contexto pode envolver a criação de definições padrão para termos e conceitos-chave.

A harmonização de termos e conceitos também é uma parte importante da gestão de dados, onde é importante que todos os sistemas de dados usem e entendam os dados da mesma maneira. Isso pode envolver a criação de um modelo de dados comum, com definições claras de todos os termos e conceitos usados.

Em resumo, a harmonização de termos e conceitos é sobre garantir que todos estejam "falando a mesma língua" para evitar confusão e melhorar a eficiência e a precisão.

Fonte: OpenAI (2024).

Quadro 31 - Resposta *prompt* 1 - GPT-4 temperatura 1

A harmonização de termos e conceitos é um processo que busca unificar a compreensão e a interpretação de palavras ou frases em um determinado contexto, geralmente em contextos técnicos, acadêmicos ou profissionais. Em outras palavras, é a prática de estabelecer consistência e uniformidade nos termos e conceitos usados.

Por exemplo, em uma empresa que opera em vários países, diferentes equipes podem usar termos diferentes para se referir à mesma coisa. Isso pode levar à confusão e a mal-entendidos. A harmonização de termos nesse contexto significa que todos na empresa concordam em usar os mesmos termos para evitar confusão.

<sup>65</sup> Na seção 5.3.1 Terminologia e LLMs: Instruções e estratégias de *prompt* para a tarefa de harmonização terminológica, abordamos os parâmetros definidos para a automatização da tarefa de harmonização. A temperatura é um dos parâmetros que podemos trabalhar em um LLM. Desse modo, a temperatura está relacionada com a criatividade da resposta fornecida pelo modelo de linguagem.

Outro exemplo pode ser encontrado no campo da saúde, onde a harmonização pode envolver a padronização de terminologia médica e de saúde para garantir que todos os profissionais de saúde entendam e usem os mesmos termos e conceitos.

No âmbito acadêmico, especialmente em disciplinas como filosofia e sociologia, a harmonização de conceitos é fundamental para garantir que todos os acadêmicos e pesquisadores tenham o mesmo entendimento sobre o significado de certos termos-chave e conceitos.

Fonte: OpenAI (2024).

Podemos observar que, de forma geral, o modelo, nas diferentes temperaturas, apresenta uma definição sobre harmonização de termos e conceitos, além de trazer alguns exemplos para ilustrar a tarefa. Percebe-se que o GPT-4 32k em temperatura 0 apresenta mais exemplos, como em “diferentes pessoas, grupos ou sistemas usem e entendam termos e conceitos da mesma maneira. Isso é especialmente importante em áreas como a ciência da informação, a gestão de dados, a pesquisa e o desenvolvimento, onde a comunicação clara e precisa é essencial.” e em “Um departamento pode se referir a "clientes" enquanto outro usa o termo "contas". Isso pode levar a confusão e ineficiência.”. Além disso, apresenta mais esclarecimentos sobre a tarefa e a sua finalidade, como, por exemplo, “A harmonização de termos e conceitos envolveria a criação de um vocabulário comum que todos os departamentos usariam.”, “A harmonização nesse contexto pode envolver a criação de definições padrão para termos e conceitos-chave.” e em “A harmonização de termos e conceitos também é uma parte importante da gestão de dados, onde é importante que todos os sistemas de dados usem e entendam os dados da mesma maneira. Isso pode envolver a criação de um modelo de dados comum, com definições claras de todos os termos e conceitos usados.”. Em relação à linguagem utilizada pelo modelo, atenta-se para a forma como as informações são tratadas. Enquanto o modelo em temperatura 0 descreve a tarefa de forma mais técnica, com exemplos claros e mais informações práticas, o modelo em temperatura 1 utiliza uma linguagem mais ‘informal’ e sem exemplos especializados. De modo geral, ambas as respostas do modelo nas diferentes temperaturas apresentam uma definição da tarefa da harmonização, ainda que o modelo em temperatura 0 traga mais dados. Vale enfatizar que, pelas descrições da harmonização apresentadas pelo modelo, podemos perceber que a base da harmonização do modelo é a abordagem desenvolvida pela ISO, visto que só considera as definições dos termos para essa atividade.

Para o segundo *prompt* – Considerando a harmonização de termos e conceitos, como eu poderia harmonizar os seguintes termos: empregado, colaborador, funcionário, trabalhador temporário, recurso externo, usuário de colaboração, operário, trabalhador e

membro do pessoal. Qual seria o termo principal e quais seriam as variantes? –, instruímos o modelo a harmonizar o conjunto de termos, sem a adição de exemplos e definições, e obtivemos as seguintes respostas no Quadros 32 e 33.

Quadro 32 - Resultado *prompt* 2 - GPT-4 32k temperatura 0

A harmonização de termos e conceitos é um processo que visa a padronização da linguagem para evitar confusões e mal-entendidos. No contexto que você apresentou, todos os termos estão relacionados a indivíduos que desempenham algum tipo de trabalho ou função dentro de uma organização.

O termo principal poderia ser "Funcionário", pois é um termo amplamente utilizado e compreendido que pode abranger todos os outros termos. No entanto, a escolha do termo principal pode depender do contexto específico e da cultura da organização.

As variantes poderiam ser definidas da seguinte forma:

1. Empregado: Este termo é frequentemente usado para se referir a alguém que é contratado por uma empresa em um sentido mais formal. Pode ser usado como sinônimo de "funcionário", mas também pode ter implicações legais específicas em alguns contextos.

2. Colaborador: Este termo é frequentemente usado para enfatizar a contribuição individual para o trabalho em equipe ou para a organização como um todo. Pode ser usado como sinônimo de "funcionário", mas também pode ser usado para incluir pessoas que não são empregadas diretamente pela organização, como consultores ou freelancers.

3. Trabalhador Temporário: Este termo é usado para se referir a alguém que é contratado por um período de tempo específico, geralmente para um projeto ou tarefa específica.

4. Recurso Externo: Este termo é usado para se referir a alguém que não é empregado diretamente pela organização, mas que fornece algum tipo de serviço ou expertise. Isso pode incluir consultores, freelancers, ou empresas terceirizadas.

5. Usuário de Colaboração: Este termo parece se referir a alguém que colabora com a organização de alguma forma, mas não é necessariamente um empregado. Pode incluir parceiros de negócios, clientes, ou outros stakeholders.

6. Operário: Este termo é frequentemente usado para se referir a trabalhadores manuais ou de produção, especialmente em um contexto industrial.

7. Trabalhador: Este é um termo muito geral que pode se referir a qualquer pessoa que realiza trabalho. Pode ser usado como sinônimo de "funcionário", mas também pode incluir pessoas que não são empregadas diretamente pela organização.

8. Membro do Pessoal: Este termo é semelhante a "funcionário", mas pode ser usado para enfatizar a inclusão de alguém na equipe ou na organização.

Lembre-se, a harmonização de termos e conceitos deve levar em conta o contexto específico e a cultura da organização. É importante garantir que todos na organização entendam e usem os termos da mesma maneira.

Fonte: OpenAI (2024).

Como podemos observar no Quadro 32, o modelo GPT-4 32k com temperatura 0 apresenta 'funcionário' como termo principal, justificando que ele pode englobar todos os outros termos por ser uma unidade mais ampla. Em seguida, como variantes, o modelo apresenta: empregado – visto que ele é sinônimo de funcionário, mas está relacionado com questões legais a depender do contexto –; colaborador – também sinônimo de funcionário, mas engloba não somente os funcionários permanentes da

empresa, como também os funcionários externos –; trabalhador temporário – referindo-se ao funcionário que exerce uma função por um tempo determinado –; recurso externo – remetendo ao funcionário terceirizado ou *freelancer*, que desenvolve tarefas de uma determinada *expertise* –; usuário de colaboração – que abrange pessoas que colaboram com a empresa, mas que não, necessariamente, são empregados, como acionistas –; operário – para denotar aos trabalhadores manuais –; trabalhador – para tratar das pessoas que são funcionárias e que também incluem pessoas que não são empregadas diretamente pela empresa –; e, finalmente, membro do pessoal – para se referir ao funcionário e que enfatiza a inclusão de alguém da empresa. Vale destacar que o modelo salienta que os resultados gerados por ele podem variar de acordo com o contexto e o domínio, além de ser necessário considerar a estruturação da empresa na qual será implantada a harmonização.

#### Quadro 33 - Resultado *prompt* 2 - GPT-4 32k temperatura 1

A harmonização terminológica é um processo que ajuda a garantir a consistência e o entendimento comum em um determinado campo. No contexto da relação de trabalho que você apresentou, é possível considerar "Funcionário" como o termo principal por ser amplamente usado e reconhecido em muitas regiões e culturas. O termo "funcionário" pode encapsular uma variedade de funções e tipos de trabalho, e muitas vezes é usado de forma genérica para se referir a qualquer pessoa que trabalhe para uma organização.

As variantes poderiam ser:

1. "Empregado": Este é um sinônimo comum de funcionário e é frequentemente usado em contextos legais e contratuais.
  2. "Colaborador": Esta é uma forma mais moderna e cooperativa de se referir a um funcionário, sugerindo um relacionamento de trabalho mais igualitário.
  3. "Trabalhador Temporário": Este termo refere-se a um funcionário que tem um contrato de trabalho temporário, em vez de um contrato de trabalho permanente.
  4. "Recurso Externo": Este termo é usado geralmente para se referir a um funcionário terceirizado ou consultor.
  5. "Usuário de Colaboração": Este termo não é comumente usado para se referir a um funcionário, mas pode ser usado em contextos onde um indivíduo está colaborando com uma equipe ou organização sem ser um funcionário oficial.
  6. "Operário": Este termo geralmente se refere a trabalhadores manuais ou operacionais.
  7. "Trabalhador": Este é um termo genérico que pode se referir a qualquer pessoa que realiza trabalho, independente do tipo de contrato de trabalho.
  8. "Membro do Pessoal": Este termo se refere a um indivíduo que é parte de uma equipe ou organização, geralmente em um sentido mais formal ou institucional.
- É importante ressaltar que a harmonização terminológica pode variar dependendo do contexto específico, da região, do setor, etc. Essa decisão deve sempre levar em conta o público-alvo e o propósito da comunicação.

Fonte: OpenAI (2024).

O modelo em temperatura 1 apresenta a mesma harmonização dos termos e conceitos: (i) funcionário como termo principal, (ii) empregado, (iii) colaborador, (iv)

trabalhador temporário, (v) recurso externo, (vi) usuário de colaboração, (vii) operário, (viii) trabalhador e (ix) membro do pessoal como variantes. É interessante observar a forma como o modelo em temperatura 1 apresenta as informações. Para a harmonização do conjunto de termos, não traz muitas informações técnicas e explicações, colocando, no máximo, dois exemplos de tipos de trabalhos. Se compararmos a disposição das informações com os resultados do modelo em temperatura 0, podemos notar que o primeiro traz mais contexto e relações com a instituição e os papéis desempenhados pelo empregado, ao passo que, em temperatura 1, as informações são as mesmas, mas não tão desenvolvidas.

Para o terceiro *prompt*, adicionamos as definições dos termos retiradas da base terminológica, sem os exemplos. Considerando isso, obtivemos as seguintes harmonizações (Quadros 34 e 35).

#### Quadro 34 - Resultado *prompt* 3 - GPT-4 32k temperatura 0

Com base nas definições fornecidas, parece que todos esses termos se referem a indivíduos que contribuem para a criação de bens ou serviços para uma empresa, mas diferem em termos de seu status de emprego, suas responsabilidades e seu relacionamento com a empresa. O termo principal poderia ser "Colaborador", pois é um termo amplo que pode abranger todos os outros termos. Ele se refere a qualquer pessoa que contribui para a empresa, independentemente do tipo de contrato de trabalho ou do status de emprego.

As variantes poderiam ser definidas da seguinte forma:

1. "Empregado" e "Funcionário": Estes são colaboradores que têm um contrato de trabalho formal com a empresa. Eles estão vinculados às políticas e regulamentos da empresa e podem ser internos ou externos.
2. "Trabalhador Temporário": Este é um colaborador que não tem um contrato de emprego permanente ou de longo prazo. Eles trabalham para a empresa em uma base não permanente.
3. "Recurso Externo": Este é um colaborador que é contratado como funcionário externo. Eles podem ser terceirizados e não estão necessariamente vinculados às políticas e regulamentos da empresa.
4. "Usuário de Colaboração": Este é um colaborador que participa de um cenário de colaboração. Eles podem ser funcionários terceirizados ou trabalhadores temporários.
5. "Operário": Este é um colaborador que pode operar um dispositivo ou trabalhar sem um dispositivo. Este termo pode se referir a trabalhadores manuais ou operários.
6. "Trabalhador" e "Membro do Pessoal": Estes são termos genéricos que podem se referir a qualquer tipo de colaborador, independentemente do status de emprego ou das responsabilidades.

Por favor, note que a harmonização de termos pode variar dependendo do contexto e das necessidades específicas da empresa.

Fonte: OpenAI (2024).

#### Quadro 35 - Resultado *prompt* 3 - GPT-4 32k temperatura 1

Com base nas definições acima, o termo principal seria "Funcionário". Isso porque todas as definições apontam para um indivíduo que realiza trabalho ou contribui para uma empresa ou organização de alguma forma, seja a longo prazo, temporária, interna ou externamente.

As variantes seriam as seguintes:

1. "Empregado": Uma subcategoria de funcionário que trabalha sob um contrato de trabalho.
2. "Colaborador": Esta palavra tem um significado amplo e geralmente se refere a qualquer pessoa que contribui para a empresa. Isso pode incluir tanto funcionários internos como externos.
3. "Funcionário Interno": Este é um funcionário que tem uma relação mais formal e direta com a empresa, geralmente sob a forma de um contrato de trabalho.
4. "Funcionário Externo": Esta é uma pessoa que contribui para a empresa, mas não como um funcionário direto.
5. "Trabalhador Temporário": Este é um tipo de funcionário que trabalha para a empresa em uma base não-permanente.
6. "Recurso Externo": Esta também é uma pessoa que contribui para a empresa, mas em um contexto mais amplo. Pode ser um contratado, um consultor, ou mesmo uma entidade como uma outra empresa.
7. "Usuário de Colaboração": Este termo se refere a um funcionário terceirizado que trabalha em parceria com a empresa.
8. "Operário": Um tipo específico de funcionário, geralmente envolvido em trabalho manual ou trabalho que envolve a operação de maquinário.
9. "Trabalhador": Este termo é bastante amplo e pode se referir a qualquer um que realiza trabalho. Dentro do contexto das definições acima, pode se referir a qualquer tipo de funcionário.
10. "Membro do Pessoal": Este é semelhante a um funcionário, mas o termo geralmente implica um nível de responsabilidade ou status mais elevados.

Fonte: OpenAI (2024).

Analisando os resultados do modelo nas distintas temperaturas, podemos notar que há diferenças quanto à harmonização dos termos. O modelo em temperatura 0 apresenta o termo principal e as variantes, indicando uma mudança em relação à primeira tentativa de harmonização. Por sua vez, o modelo em temperatura 1 traz uma harmonização mantendo o termo principal da primeira tentativa; no entanto, adiciona novos termos nessa etapa.

No que tange à harmonização em temperatura 0, o termo principal é ‘colaborador’, e as variantes são ‘empregado’ e ‘funcionário’ como variantes intercambiáveis, referindo-se tanto ao trabalhador interno como ao externo. As outras variantes – ‘trabalhador temporário’, ‘recurso externo’, ‘usuário de colaboração’ e ‘operário’ - seguem a mesma ordem da primeira harmonização. Por fim, o modelo seleciona ‘trabalhador’ e ‘membro do pessoal’ como variantes intercambiáveis, uma vez que são termos mais genéricos e que abrangem qualquer tipo de colaborador. Vale evidenciar que o modelo apresentou, de início, uma terminologia mais abrangente, vide a seleção de ‘colaborador’ como termo principal e ‘empregado’ e ‘funcionário’ como primeiras variantes. Ao final, selecionou ‘trabalhador’ e ‘membro do pessoal’, também termos mais amplos, como variantes, considerando que eles podem abarcar pessoas internas e externas no contexto da instituição. Embora a ordem das variantes não seja

um fator a ser avaliado aqui, posto que, ao digitar um dos termos harmonizados, ele será substituído automaticamente pelo termo principal, ainda é válido compreender as justificativas que o modelo fornece sobre a seleção dos termos.

A harmonização feita pelo modelo em temperatura 1 apresentou uma seleção diferente, trazendo adições de termos, como, por exemplo, ‘funcionário interno’ e ‘funcionário externo’. A partir da instrução, observou-se que ‘funcionário’ é o termo principal, seguido de ‘empregado’, ‘colaborador’, ‘funcionário interno’, ‘funcionário externo’, ‘trabalhador temporário’, ‘recurso externo’, ‘usuário de colaboração’, ‘operário’, ‘trabalhador’ e ‘membro do pessoal’ como variantes. Acerca das adições, podemos notar que ‘funcionário interno’ se refere ao profissional que tem uma relação formal (contratual) com a empresa, o que, de certo modo, liga-se com o termo ‘empregado’, o qual também é um funcionário que tem vínculo empregatício com uma empresa e segue um contrato de trabalho. O termo ‘funcionário externo’, considerando as definições apresentadas pelo LLM, aproxima-se de ‘trabalhador temporário’, ‘recurso externo’ e ‘usuário de colaboração’. Nesse sentido, todos os termos se referem à pessoa que não trabalha sob um vínculo empregatício direto com a empresa e nem de forma permanente. Pensando nisso, as inserções terminológicas feitas pelo modelo não são tão necessárias para essa harmonização.

Quanto aos resultados obtidos a partir do quarto *prompt* – definições com exemplos –, podemos observar a omissão do termo ‘empregado’ no modelo em temperatura 0 e uma nova harmonização do modelo em temperatura 1, conforme os Quadros 36 e 37.

#### Quadro 36 - Resultado *prompt* 4 - GPT-4 32k temperatura 0

Com base nas definições fornecidas, parece que o termo principal poderia ser "Funcionário". Este termo é amplo o suficiente para abranger todos os tipos de trabalhadores que contribuem para a criação de bens ou serviços para uma empresa, independentemente do tipo de contrato de trabalho ou da duração do emprego.

Aqui estão as variantes:

1. "Colaborador": Este termo pode ser usado para se referir a qualquer pessoa que contribua para a empresa, seja um funcionário interno ou externo.
2. "Trabalhador Temporário": Este é um tipo específico de funcionário que não tem um contrato de trabalho permanente ou de longo prazo.
3. "Recurso Externo": Este termo pode ser usado para se referir a um funcionário que é contratado como funcionário externo, como consultores ou auditores externos.
4. "Usuário de Colaboração": Este é um tipo específico de funcionário terceirizado que participa de um cenário de colaboração, como o gerenciamento de inventário.
5. "Operário": Este termo pode ser usado para se referir a um funcionário que opera um dispositivo ou trabalha sem um dispositivo, como um motorista de empilhadeira ou um

montador no chão de fábrica.

6. "Trabalhador": Este é um termo genérico que pode ser usado para se referir a qualquer pessoa que trabalha para a empresa, independentemente do tipo de contrato de trabalho ou da duração do emprego.

7. "Membro do Pessoal": Este termo pode ser usado para se referir a qualquer pessoa que faz parte da equipe de uma empresa, independentemente do tipo de contrato de trabalho ou da duração do emprego.

Lembre-se de que a harmonização de termos depende muito do contexto específico e das convenções de terminologia usadas em sua organização ou indústria. Portanto, essas sugestões podem precisar ser ajustadas para se adequar ao seu contexto específico.

Fonte: OpenAI (2024).

#### Quadro 37 - Resultado *prompt* 4 - GPT-4 32k temperatura 1

De acordo com as definições fornecidas, o termo "empregado" poderia ser considerado o termo principal. O termo "empregado" parece abranger todos os indivíduos que contribuem para a criação de bens ou serviços para uma empresa, independentemente da natureza exata de seu contrato de trabalho ou sua relação específica com a empresa.

Os demais termos poderiam ser considerados variantes do termo "empregado", cada um representando diferentes tipos ou categorias de empregados, com base em fatores como a duração do seu contrato de trabalho, se eles são internos ou externos, a natureza específica do seu trabalho, etc.

Aqui está uma possível harmonização dos termos:

##### 1. Empregado (termo principal)

- Colaborador: Um tipo de empregado que pode ter um contrato de trabalho ou contrato de serviços, seja interno ou externo.

- Funcionário: Um tipo de empregado interno que está vinculado às instruções e deve aderir às políticas e regulamentos da empresa.

- Trabalhador temporário: Um tipo de empregado que não tem um contrato explícito de emprego permanente ou de longo prazo.

- Recurso externo: Um tipo de empregado, frequentemente um consultor ou auditor externo, que é terceirizado e compartilha o endereço de trabalho do empregador contratado por um determinado período de tempo.

- Usuário de colaboração: Um tipo de empregado terceirizado que participa de um cenário de colaboração.

- Operário: Um tipo de empregado que pode operar um dispositivo ou trabalhar sem um dispositivo.

- Trabalhador: Um termo mais genérico que pode ser usado como sinônimo de empregado.

- Membro do pessoal: Outro termo genérico que pode ser usado como sinônimo de empregado.

Cabe ressaltar que a harmonização dos termos pode variar dependendo do contexto específico e das políticas e práticas específicas de uma empresa ou setor.

Fonte: OpenAI (2024).

No que diz respeito à harmonização do GPT-4 32k em temperatura 0, percebemos que o termo principal é 'funcionário', e as variantes são 'colaborador', 'trabalhador temporário', 'recurso externo', 'usuário de colaboração', 'operário', 'trabalhador' e 'membro do pessoal'. Como já dito, com o uso dessa estratégia, o modelo optou por omitir o termo 'empregado'. Isso pode ter ocorrido devido à aproximação semântica entre os termos 'funcionário', 'colaborador' e 'empregado'.



Contudo, é importante atentar para as questões legais intrínsecas ao termo ‘empregado’, visto que, na legislação brasileira, utiliza-se empregado para definir a pessoa que trabalha para uma empresa/instituição.

A harmonização elaborada pelo modelo em temperatura 1, de outra parte, selecionou como termo principal ‘empregado’, seguido das variantes ‘colaborador’, ‘funcionário’, ‘trabalhador temporário’, ‘recurso externo’, ‘usuário de colaboração’, ‘operário’, ‘trabalhador’ e ‘membro do pessoal’. Essa proposta de harmonização se assemelha mais ao resultado da aplicação da tarefa feita manualmente, visto que seleciona como termo principal ‘empregado’, considerando toda a abrangência do termo. Essa nova seleção de termos indica que a estratégia de *prompt* utilizada foi bem-sucedida, já que a harmonização, em certa medida, aproxima-se da nossa.

Esses resultados evidenciam dois pontos interessantes em relação aos LLMs. O primeiro se refere à adição de informações que aprimoram os resultados do modelo. Nos quadros anteriores, podemos perceber que, ao inserirmos mais camadas de informação, como no uso da última estratégia de *prompt*, obtivemos um resultado muito próximo da harmonização desenvolvida manualmente. Além disso, podemos observar que o ajuste de temperatura também é uma variável a ser refletida. Enquanto o modelo em temperatura 0 não se aproximou da nossa proposta de harmonização, o modelo em temperatura 1 apresentou um resultado mais semelhante ao nosso. Vale destacar que a temperatura está relacionada com a criatividade e liberdade do modelo para gerar sua resposta. Portanto, a temperatura mais elevada permitiu que o LLM apresentasse uma nova proposta, ainda que, no terceiro *prompt*, vimos que ele adicionou termos que não eram necessários nessa harmonização.

Após realizarmos os testes fornecendo apenas informações de definições terminológicas e exemplos, passamos a adicionar outra camada de informação gerada da verbalização do nosso *Knowledge Graph*. No Quadro 38, podemos examinar os resultados obtidos a partir da representação do conhecimento.

Quadro 38 - Resultado *prompt* 5 - GPT-4 32k temperatura 0

A harmonização de termos é um processo que visa a padronização e a consistência na utilização de termos dentro de um determinado contexto. No caso apresentado, os termos estão relacionados a diferentes tipos de trabalhadores ou colaboradores dentro de um sistema ERP.

Aqui está uma sugestão de como você pode harmonizar esses termos:

1. **\*\*Termo Principal: Empregado\*\***

- Variantes: Colaborador, Funcionário, Membro do Pessoal

2. **\*\*Termo Principal: Trabalhador Temporário\*\***  
- Variantes: Nenhuma (este termo é bastante específico e não parece ter uma variante direta nos termos fornecidos)
  3. **\*\*Termo Principal: Recurso Externo\*\***  
- Variantes: Nenhuma (este termo é bastante específico e não parece ter uma variante direta nos termos fornecidos)
  4. **\*\*Termo Principal: Usuário de Colaboração\*\***  
- Variantes: Nenhuma (este termo é bastante específico e não parece ter uma variante direta nos termos fornecidos)
  5. **\*\*Termo Principal: Operário\*\***  
- Variantes: Trabalhador (este termo pode ser usado de forma mais geral, mas pode ser considerado uma variante de "Operário" em contextos onde o trabalho manual ou físico é o foco)
- Note que a harmonização de termos pode variar dependendo do contexto específico e das necessidades da organização. Além disso, pode haver situações em que um termo não tem uma variante direta ou é usado de forma única em um determinado contexto.
- Por fim, é importante lembrar que a harmonização de termos é um processo contínuo que pode exigir revisões e ajustes à medida que novos termos são introduzidos ou as necessidades da organização mudam.

Fonte: OpenAI (2024).

Como mostra o Quadro 38, o modelo GPT-4 32k em temperatura 0 apresenta uma estrutura de harmonização completamente diferente das propostas anteriores. Nesse caso, o modelo gerou um termo principal e as variantes para cada termo, ou seja, não sugeriu um termo único como termo principal e o restante como variantes. O modelo propôs cinco termos principais e suas variantes, como podemos ver na sistematização do Quadro 39.

Quadro 39 - Termos principais e variantes propostos pelo modelo GPT-4 32k

<b>Termo principal</b>	<b>Variantes</b>
<b>Empregado</b>	Colaborador Funcionário Membro do pessoal
<b>Trabalhador temporário</b>	Nenhuma variante
<b>Recurso externo</b>	Nenhuma variante
<b>Usuário de colaboração</b>	Nenhuma variante
<b>Operário</b>	Trabalhador

Fonte: Elaborado pela autora.

No quadro anterior, podemos observar uma estrutura de harmonização distinta das fornecidas previamente e não segue a nossa instrução de apresentar apenas um termo principal e as demais serem variantes. No entanto, é interessante analisar o processo de harmonização proposto pelo modelo, que reúne termos semanticamente semelhantes. No grupo em que ‘empregado’ é escolhido como termo principal, e ‘colaborador’, ‘funcionário’ e ‘membro do pessoal’ como variantes, podemos observar que há uma aproximação entre os termos, uma vez que todos se referem ao indivíduo

que trabalha para uma instituição/organização e são unidades mais abrangentes, possibilitando englobar todos os tipos de empregados.

Para além disso, vale refletir sobre o fato de os termos ‘trabalhador temporário’, ‘recurso externo’ e ‘usuário de colaboração’ serem termos principais e não apresentarem variantes. Se considerarmos que tais unidades lexicais são muito específicas, certamente não apresentariam variações. A justificativa do modelo, para esses casos, é que outros termos que poderiam ser variantes não foram fornecidos. Tendo em vista que o nosso objetivo é selecionar um termo principal dentre as unidades lexicais do conjunto e as demais serem variantes, essa tentativa de harmonização gerada pelo modelo não funciona, já que ele não soluciona o problema geral da tarefa, pois, a princípio, deveria seguir a linha de raciocínio de escolher o termo mais abrangente como principal e o restante ser definido como variações.

O modelo GPT-4 32k em temperatura 1, por sua vez, apresentou uma estrutura de harmonização condizente com as estruturas anteriores, como podemos visualizar no Quadro 40.

Quadro 40 - Resultado *prompt* 5 - GPT-4 32k temperatura 1

A harmonização dos termos em um ambiente de ERP é crucial para manter a consistência e entender corretamente as funções e responsabilidades associadas a cada papel. Dito isso, é importante notar que o significado exato de cada termo pode variar dependendo do contexto específico e da cultura organizacional. Portanto, a seguinte proposta é um exemplo geral e pode ser necessário ajustá-la de acordo com as necessidades específicas da sua empresa.

Termo Principal: "Empregado"

Variantes:

1. Colaborador: Este é geralmente um sinônimo de empregado, mas pode sugerir um envolvimento ou contribuição mais ativa para o trabalho em equipe e a cultura organizacional.
2. Funcionário: Este termo é muito semelhante a empregado e colaborador, embora possa implicar um nível de senioridade ou profissionalismo um pouco mais elevado em alguns contextos.
3. Trabalhador Temporário (ou Contingent Worker): Esta é uma subcategoria de empregado que especificamente se refere a indivíduos que são contratados por um período de tempo limitado.
4. Recurso Externo: Este termo é frequentemente usado para se referir a empregados temporários, freelancers ou consultores, que não são empregados permanentes, mas que contribuem para a organização.
5. Usuário de Colaboração: Embora possa se referir a um empregado em alguns contextos, este termo também pode se referir a indivíduos fora da organização que têm acesso ao sistema ERP para fins de colaboração.
6. Operário: Este termo é geralmente usado para se referir a empregados que realizam trabalho manual ou físico. No contexto de um sistema ERP, no entanto, pode ser usado para se referir a qualquer usuário que realiza tarefas dentro do sistema.
7. Trabalhador: Este é um termo genérico que pode se referir a qualquer pessoa que realiza

trabalho, seja um empregado permanente, temporário ou um consultor.  
 8. Membro do Pessoal: Este termo é semelhante a empregado, mas pode sugerir um sentido de pertencimento e contribuição a longo prazo para a organização.  
 Lembre-se que essas são apenas sugestões baseadas em usos comuns desses termos e o significado exato e as implicações de cada um podem variar dependendo do contexto específico e da cultura da sua empresa. Sempre é aconselhável estabelecer definições claras e consistentes para esses termos em sua documentação organizacional para evitar confusão.

Fonte: OpenAI (2024).

Com essa estratégia de *prompt*, o modelo em temperatura 1 selecionou ‘empregado’ como termo principal (escolha que se assemelha a nossa harmonização) e ‘colaborador’, ‘funcionário’, ‘trabalhador temporário’, ‘recurso externo’, ‘usuário de colaboração’, ‘operário’, ‘trabalhador’ e ‘membro do pessoal’ como variantes. Se compararmos essa harmonização com a desenvolvida pelo modelo no *prompt* 4, podemos perceber que não há nenhuma variação, isto é, ele manteve a mesma ordem dos termos. É válido considerar, nessa harmonização, as justificativas do modelo para a sua seleção. Podemos notar que ele mistura termos mais abrangentes com os termos mais específicos, não mantendo uma ordem do termo mais amplo para o mais específico. Isso pode ser visto na justificativa dos termos ‘trabalhador’ e ‘membro do pessoal’, os quais são definidos pelo modelo como termos mais genéricos que englobam os mais variados tipos de empregados.

Como dito anteriormente, os resultados obtidos com o *prompt* 6 – inclusão de informações das definições da base terminológica, dos exemplos e da verbalização do *Knowledge Graph* –, foram do modelo GPT-4 nas temperaturas 0 e 1. O modelo GPT-4 32k não conseguiu processar todas as informações dadas e, por isso, optamos por avaliar o *prompt* em um outro modelo. Nos quadros 41 e 43, podemos analisar as respostas geradas pelo modelo nessas condições.

#### Quadro 41 - Resultado *prompt* 6 - GPT-4 temperatura 0

A harmonização de termos em um sistema ERP pode ser um desafio, especialmente quando se lida com diferentes contextos e nuances de significado. No entanto, com base nas definições e verbalizações fornecidas, aqui está uma sugestão de como você pode harmonizar esses termos:

Termo principal: Empregado

- Variantes: Colaborador, Funcionário, Membro do pessoal

Esses termos parecem ser usados de forma intercambiável e se referem a uma pessoa que contribui para a criação de bens e serviços em uma empresa sob um contrato de trabalho.

Termo principal: Trabalhador Temporário

- Variantes: Nenhuma

Este termo parece ser usado para se referir a profissionais independentes, trabalhadores com contrato temporário, empreiteiros ou consultores.

Termo principal: Recurso Externo  
 - Variantes: Nenhuma  
 Este termo parece ser usado para se referir a um funcionário terceirizado contratado como funcionário externo, como consultores ou auditores externos.

Termo principal: Usuário de Colaboração  
 - Variantes: Nenhuma  
 Este termo parece ser usado para se referir a um funcionário terceirizado que participa de um cenário de colaboração.

Termo principal: Operário  
 - Variantes: Trabalhador  
 Estes termos parecem ser usados para se referir a um indivíduo que opera um dispositivo ou trabalha sem um dispositivo.

Por favor, note que a harmonização de termos pode variar dependendo do contexto específico e das necessidades da sua organização. Além disso, é importante garantir que todos na organização entendam e usem os termos harmonizados de forma consistente para evitar confusão.

Fonte: OpenAI (2024).

Seguindo a estrutura de harmonização sugerida pelo modelo GPT-4 32k em temperatura 0, o modelo GPT-4 (temperatura 0) também apresentou um termo principal e variantes para cada unidade lexical. Dessa forma, podemos observar cinco harmonizações, conforme Quadro 42.

Quadro 42 - Termos principais e variantes propostos pelo modelo GPT-4

<b>Termo principal</b>	<b>Variantes</b>
<b>Empregado</b>	Colaborador Funcionário Membro do pessoal
<b>Trabalhador temporário</b>	Nenhuma variante
<b>Recurso externo</b>	Nenhuma variante
<b>Usuário de colaboração</b>	Nenhuma variante
<b>Operário</b>	Trabalhador

Fonte: Elaborado pela autora.

Como discutimos previamente, a estrutura de harmonização do quadro 42 não segue a abordagem metodológica desenvolvida neste trabalho, ou seja, não apresenta um termo principal selecionado dentre as unidades e o restante como variações. Tal harmonização gerada pelo modelo não resolve o problema, visto que não fornece uma harmonização para todo o conjunto. Se considerarmos o que foi apresentado inicialmente pelo modelo no *prompt* 1, – o que é harmonização de termos e conceitos –, é perceptível que a abordagem selecionada aqui não contribui para a tarefa.

No que se refere à resposta gerada pelo modelo em temperatura 1, podemos notar que a estrutura da abordagem de harmonização é mantida, isto é, o modelo selecionou um termo principal dentre as unidades do conjunto, e o restante são

variantes. Além disso, podemos observar algumas variações com relação à ordem dos termos dessa harmonização para a anterior.

#### Quadro 43 - Resultado *prompt* 6 - GPT-4 temperatura 1

Os termos que você listou são, em muitos casos, intercambiáveis dependendo do contexto, e alguns têm implicações sutis diferentes. No entanto, considerando as definições e a verbalização fornecida, os termos podem ser harmonizados da seguinte maneira:

- **Termo principal: Empregado**. Este termo é bastante genérico e pode se referir a qualquer pessoa que contribua ou contribuiu para a criação de bens e serviços na empresa, seja como interno ou externo. Sob esse guarda-chuva, cairiam todas as outras variações que se referem a tipos específicos de empregados.
- **Variantes**:
  - **Funcionário**: Pode ser considerado sinônimo de empregado, mas implica uma relação mais formal e estável. Este termo é frequentemente usado para se referir a pessoas que têm um contrato de trabalho de longo prazo com uma empresa.
  - **Trabalhador Temporário/Contingente**: Esta variante especifica um tipo de empregado que não tem um contrato de longo prazo com a empresa. Eles podem estar lá por um período de tempo definido (temporário) ou para cobrir as necessidades da empresa em um determinado momento (contingente).
  - **Recurso Externo**: Esta variante se refere a uma pessoa que é tecnicamente empregada por outra entidade (como uma empresa de consultoria ou uma agência de empregos), mas que é contratada para trabalhar para a empresa por um tempo definido.
  - **Usuário de Colaboração**: Este termo se refere a pessoas que não são empregados em tempo integral, mas que trabalham em conjunto com a empresa em certos projetos ou iniciativas. Eles podem ser funcionários de outra empresa ou trabalhadores independentes.
  - **Operário/Trabalhador**: Esses termos são frequentemente usados para se referir a empregados que realizam trabalho manual ou operacional, como operadores de máquinas em uma linha de produção.
  - **Colaborador**: Este é um termo bastante abrangente que pode se referir a qualquer pessoa que trabalhe com a empresa, seja como empregado, trabalhador temporário, recurso externo, etc.
  - **Membro do Pessoal**: Este termo normalmente se refere a pessoas que têm uma relação de longo prazo com a empresa e desempenham funções administrativas ou de gestão. No entanto, vale lembrar que a nomenclatura exata e as definições podem variar dependendo das práticas e normas da empresa. Além disso, as leis trabalhistas do país onde a empresa opera também podem influenciar essas definições.

Fonte: OpenAI (2024).

Nessa harmonização, o modelo em temperatura 1 apresentou o termo ‘empregado’ como o principal e ‘funcionário’, ‘trabalhador temporário’, ‘recurso externo’, ‘usuário de colaboração’, ‘operário/trabalhador’, ‘colaborador’ e ‘membro do pessoal’ como variantes. Para essa última instrução, o modelo apresenta os termos ‘operário’ e ‘trabalhador’ como iguais, justificando que ambos se referem aos empregados que desenvolvem trabalhos manuais ou operacionais. Em todos os resultados obtidos, essa foi a primeira ocorrência para unir tais unidades lexicais e tratá-las como semelhantes.

No que diz respeito à ordem das variantes, nesse último resultado sugerido pelo modelo, podemos observar que ‘funcionário’ apareceu antes de ‘colaborador’, apesar de ambos os termos apresentarem uma definição muito próxima. Em comparação com o resultado anterior do modelo em temperatura 1, no *prompt* 5, o LLM apresentou ‘colaborador’ como primeira variante, seguida de ‘funcionário’, enquanto no *prompt* 6, o modelo apresentou ‘funcionário’ como primeira variante e ‘colaborador’ como sexta variante. Vale frisar, novamente, que os dois resultados também apresentam uma ordem que varia entre termos mais abrangentes e termos específicos, indicando uma unidade mais ampla como termo principal e, em suas variações, encontramos termos abrangentes e específicos misturados.

Considerando as aplicações da nossa abordagem semântico-terminológica para a harmonização de termos e conceitos tanto manualmente como automaticamente e os resultados obtidos, faz-se necessário destacar alguns pontos referentes à criação da abordagem, à construção da arquitetura do *Knowledge Graph* com *frames* semânticos, à aplicação da nossa metodologia e ao uso dos LLMs para a tarefa de harmonização.

Primeiramente, quanto ao desenvolvimento da abordagem semântico-terminológica para harmonização de termos e conceitos, percebemos que a contextualização e a camada de informação linguística advinda dos *frames* semânticos fornecem subsídios para além das definições e contribuem para a tarefa de harmonizar os termos. Ao passo que a ISO propõe uma harmonização considerando apenas as definições e a relação conceitual dos termos – se um termo apresenta um ou mais conceitos relacionados a ele – sem buscar compreender e aplicar as relações conceptuais, contextuais e culturais da terminologia, nossa proposta busca adicionar uma camada extra de informação para complementar na harmonização. Esse nível de detalhamento contribui para que o responsável pela tarefa de harmonizar consiga compreender não só o termo, mas todas as relações entre ele e as variantes, bem como o contexto da terminologia, como ela é utilizada, em quais domínios ela ocorre, dentre outros tipos de dados necessários para a seleção do termo principal e das variantes. Ao considerarmos tudo o que compõe e que é inerente aos termos, passamos a ter uma visão mais ampla da situação em que determinada terminologia está inserida e, conseqüentemente, fazendo uma melhor harmonização.

Seguindo a nossa abordagem metodológica, previmos uma etapa de construção da representação do conhecimento no formato de *Knowledge Graph* com *frames* semânticos. Essa etapa foi fundamental em vários aspectos do processo de

harmonização. Em um primeiro momento, para a construção do KG, precisamos estudar de forma aprofundada todo o funcionamento do ERP e de seus componentes, passando pelos procedimentos e operações de cada recurso apresentado. Esse estudo facilita o processo de harmonização, posto que, com a aquisição do conhecimento do *software*, podemos compreender todas as partes do produto a ponto de o representarmos visualmente e de entendermos as relações entre os termos e os contextos. Dessa forma, podemos afirmar que a pesquisa feita para a elaboração das estruturas conceptuais e do KG é uma das fases mais importantes na nossa abordagem.

Outro ponto relacionado a nossa abordagem condiz com a forma de harmonizar. Se, por um lado, a proposta metodológica da ISO considera as definições dos conceitos de cada termo, podendo também levar em conta a frequência que uma dada unidade lexical apresenta, por outro lado, a nossa abordagem propõe conectar o máximo de informações que extrapolam as definições terminológicas, relacionadas aos termos para que seja possível fazer uma melhor seleção do termo principal e das variantes. Após a aplicação da nossa metodologia semântico-terminológica para harmonização de termos e conceitos, percebemos que a camada contextual fornecida pelos *frames* e pelo *Knowledge Graph* contribuiu para a tarefa. Inicialmente, tinha-se em mente que a escolha poderia ocorrer pela frequência e pela quantidade de definições que determinado termo apresenta, mas, com as informações extras advindas da representação do conhecimento, percebe-se que a localização do termo no mapa/KG permite ter uma compreensão melhor do domínio e, por conseguinte, do comportamento da terminologia, facilitando o processo de tomada de decisão. Podemos perceber que a harmonização não seria satisfatória se considerássemos apenas a proposta da ISO, já que, como vimos, é preciso trazer mais informações contextuais que, muitas vezes, não estão presentes nas definições. Na harmonização em que selecionamos ‘empregado’ como termo principal, um dos pontos para essa escolha se refere ao aspecto legal que é intrínseco a tal unidade. Ao considerarmos a legislação trabalhista brasileira, notamos que ‘empregado’ seria o termo principal pela sua relação direta com o termo ‘empregador’, fato que se reflete na estruturação dos *frames*. Além disso, trata-se de um termo mais amplo, capaz de abranger todos os outros termos, desde ‘funcionário’, que também é um termo mais genérico, até termos específicos como ‘operário’ e ‘trabalhador temporário’.

A elaboração manual da harmonização não é uma tarefa simples. Harmonizar termos e conceitos exige que se compreendam bem o contexto da harmonização e a



necessidade de executá-la. A complexidade envolvida nas etapas da tarefa torna o processo complexo, visto que são muitos estágios de estudo do procedimento, da construção de todo o método, da busca por materiais para entender a terminologia e o contexto e da aplicação da metodologia propriamente dita. Ainda que nossa abordagem apresente o *Knowledge Graph* e os *frames*, e essa camada de informações auxilia na harmonização, durante a nossa implementação da metodologia, tivemos muitas dúvidas, principalmente sobre o funcionamento de cada área do ERP. Nesse ponto, trabalhar com um especialista da área, no nosso caso da área de negócios e especialistas dos diversos setores do ERP, facilitaria ainda mais a tarefa de harmonização. Para se harmonizar termos, é imprescindível ter o conhecimento especializado do domínio e das unidades lexicais, refletindo acerca do uso de cada palavra, quando elas são aplicadas, em quais domínios elas se inserem, dentre outros questionamentos. Muitas vezes, há sutilezas que indicam qual é o termo principal e quais são as variantes e, nesses casos, o especialista que detém o conhecimento aprofundado da área poderá contribuir para a seleção.

No que se refere à harmonização feita com os modelos de linguagem, vale destacarmos alguns pontos. Primeiramente, podemos observar que o LLM tem um treinamento anterior e que sabe definir o que é a harmonização, para que serve e como é feita. Essa primeira etapa já era esperada, uma vez que os modelos recebem um treinamento durante o seu desenvolvimento e a proposta da ISO é antiga. Em segundo lugar, nas primeiras harmonizações – *prompts* 2, 3 e 4 –, a cada adição de informação (definições e exemplos), podemos perceber que o modelo conseguia fazer uma boa aplicação da tarefa. Na primeira etapa, por exemplo, o modelo, nas diferentes temperaturas, gerou o mesmo resultado. Conforme inseríamos mais dados, notamos que o desempenho dos modelos melhorou, apesar de apresentar, no caso do GPT-4 32k e GPT-4 em temperaturas 0, uma harmonização para cada termo. Ao adicionarmos a verbalização do *Knowledge Graph*, observamos que a harmonização do GPT-4 32k em temperatura 1 selecionou o termo principal ‘empregado’ e as demais variantes seguindo a mesma concepção que fizemos manualmente.

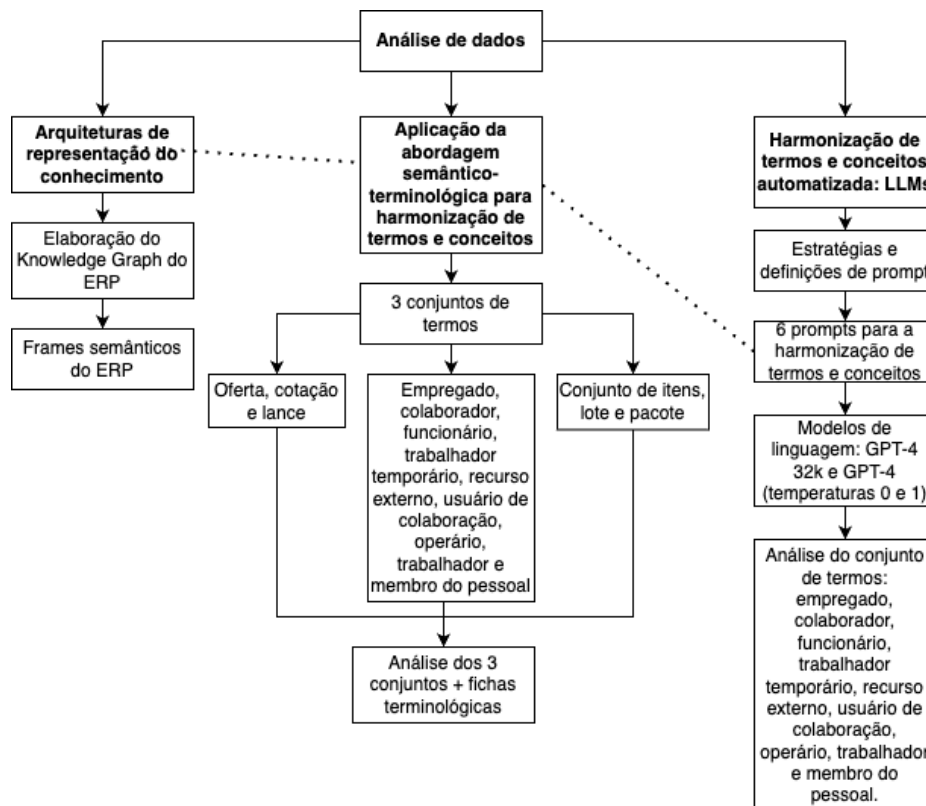
Acerca da temperatura dos modelos, constatamos que os resultados variavam conforme mudávamos a temperatura. Inicialmente, imaginávamos que o modelo em temperatura 0 apresentaria uma harmonização mais próxima da que elaboramos, já que, nessa temperatura, ele não apresenta muitas variações nem tantas alucinações. Porém, percebemos que ele tentou apresentar uma harmonização para cada termo, o que não estava descrito na instrução. Considerando os trabalhos produzidos sobre possíveis

alucinações e melhoras de performance dos modelos, acreditávamos que esses resultados seriam aprimorados com a verbalização do nosso *Knowledge Graph*. Entretanto, ele continuou apresentando uma harmonização para cada termo do conjunto.

Ainda no que se refere à temperatura dos modelos, como dito, imaginávamos que a temperatura 0 apresentaria resultados melhores. Contudo, a última harmonização do modelo em temperatura 1 apresentou soluções ótimas, exibindo uma seleção terminológica igual a nossa<sup>66</sup>. Nesse ponto, destacamos que a criatividade do modelo foi fundamental para a harmonização. Notamos que ele, nessa temperatura, considerou toda a parte contextual antes de prever o resultado. Além disso, é interessante perceber que o modelo em temperatura 1 manteve uma certa consistência de resultados, apresentando pequenas variações.

Considerando os dados analisados neste capítulo, trazemos a Figura 98 com destaque aos principais encaminhamentos da análise.

Figura 98 - Sistematização do capítulo “Análise de dados”



Fonte: Elaborada pela autora.

<sup>66</sup> Vale frisar que a ordem das variantes não é tão pertinente, uma vez que, quando a abordagem for aplicada em uma base de dados, ao se digitar uma variante, automaticamente ela será substituída pelo termo principal.

Como ilustra a Figura 98, neste capítulo, tratamos das arquiteturas de representação do conhecimento, evidenciando a elaboração do *Knowledge Graph* do ERP e dos frames semânticos do domínio. Em seguida, abordamos a aplicação da abordagem semântico-terminológica para a harmonização de termos e conceitos. Para essa aplicação manual da abordagem, selecionamos três conjuntos de dados e, ao final da análise de cada conjunto, elaboramos fichas terminológicas, indicando o termo principal, as variantes, dentre outras informações. Por fim, discorremos sobre a automatização da abordagem de harmonização de termos e conceitos de base semântico-terminológica. Nesse momento, apresentamos as seis estratégias de *prompt* selecionadas para a aplicação. Além disso, mostramos as definições dos modelos utilizados para o experimento e os parâmetros selecionados. A partir de tal etapa, escolhemos o conjunto de termos composto por ‘empregado’, ‘colaborador’, ‘funcionário’, ‘trabalhador temporário’, ‘recurso externo’, ‘usuário de colaboração’, ‘operário’, ‘trabalhador’ e ‘membro do pessoal’ para aplicarmos a harmonização e automatizá-la no LLM.

Com a realização dos experimentos, observamos que uma abordagem de base semântico-terminológica para a harmonização de termos e conceitos torna a metodologia mais robusta, visto que, ao inserir uma camada de informações semânticas provenientes dos *frames*, é possível compreender os termos e conceitos de forma mais aprofundada, considerando o contexto em que estão inseridos. No que tange à harmonização automática, percebemos que a estrutura cognitiva somada ao *Knowledge Graph* potencializou os resultados do *Large Language Model*, fornecendo uma resposta similar à harmonização manual.

Após a análise dos dados e das discussões acerca dos resultados obtidos, no próximo capítulo, apresentamos as considerações finais deste trabalho.

## 7 CONSIDERAÇÕES FINAIS

Ao longo desta pesquisa, buscamos desenvolver uma abordagem semântico-terminológica para a harmonização de termos e conceitos que abarcasse a Semântica de *Frames*, a Terminologia baseada em *Frames* (*Frame-based Terminology*) e a Inteligência Artificial. Partimos da premissa de que a Semântica de *Frames* poderia servir como princípio organizador de uma base de dados ERP (*Enterprise Resource Planning*), contribuindo para a harmonização terminológica. Além disso, também nos detemos na automatização da nossa proposta em um *Large Language Model*, utilizando diferentes estratégias de *prompt* e um *Knowledge Graph* para representar o conhecimento do *software*.

O ponto de partida para esta tese foi a necessidade de empresas que desenvolvem *softwares* para gerenciamento empresarial buscarem por modelos linguísticos para a solução de demandas de ordem tecnológica, sobretudo as questões relativas à Linguística, à Tradução e à Terminologia.

Com a transformação digital e o pleno desenvolvimento de *Large Language Models* (LLMs), como o ChatGPT, muitas atividades que, anteriormente, era realizadas manualmente estão sendo automatizadas. No contexto empresarial, antigamente, a gestão de dados de uma empresa era feita de forma manual e segmentada. Com o avanço tecnológico, desenvolveu-se um *software* capaz de gerir todos os setores de uma instituição de forma rápida e eficiente. O *Enterprise Resource Planning* (ERP) é uma solução computacional que auxilia na administração de todo um empreendimento, facilitando o compartilhamento de dados e a comunicação entre os diversos setores de uma empresa. Embora esse recurso consiga apresentar os processos e resultados das áreas corporativas, como Recursos Humanos, Finanças, Compra e Vendas, Marketing, Logística, *Supply Chain*, dentre outras, a comunicação entre os segmentos apresenta diversas variações terminológicas, causando ruídos comunicacionais, inconsistências terminológicas e divergências informacionais.

Para solucionar esse entrave terminológico, trabalha-se com a harmonização de termos e conceitos. Considerando esse contexto, nosso trabalho surgiu a partir da necessidade de dar um tratamento mais robusto para a harmonização de termos e conceitos, procedimento desenvolvido pela ISO (2007), fornecendo uma visão

contextual e cognitiva mais ampla para que essa metodologia seja mais completa. Além de o desenvolvimento de uma metodologia semântico-terminológica para a harmonização de termos e conceitos, neste trabalho, também pretendeu-se automatizá-la em um LLM, evidenciando que é possível realizar atividades de cunho terminológico em modelos de linguagem. Enquanto, no passado, a realização de tarefas voltadas para o tratamento de termos era desempenhada à mão ou com o auxílio de ferramentas computacionais para a análise e para o processamento de dados, atualmente, uma forma de automatizar os trabalhos terminológicos é por meio do uso de um LLM. Levando em conta a capacidade que um modelo de linguagem apresenta para a geração de textos em diferentes estilos e com alta fluência e precisão, percebe-se que tal recurso pode desempenhar tarefas terminológicas, como a harmonização de termos e conceitos, a partir de estratégias e técnicas de *prompt* bem elaboradas.

À vista disso, esta tese teve por objetivo, em nível teórico, investigar o papel da Semântica de *Frames* como princípio organizador de uma base de dados ERP (*Enterprise Resource Planning*), contribuindo para a harmonização terminológica. Para isso, buscamos, na Semântica de *Frames*, suporte para mapearmos os domínios do ERP e estudarmos a terminologia. Nesse sentido, a estrutura de *frames* contribuiu tanto para a representação da base de dados como para a tarefa de harmonização de termos e conceitos. Como princípio organizador, os frames auxiliaram construção da arquitetura do ERP, sobretudo quando aplicados concomitantemente com o *Knowledge Graph*. Essa representação do conhecimento se tornou mais abrangente, visto que, além de podermos visualizar a estruturação do *software*, adicionamos uma camada de descrição linguística acerca do ERP, facilitando a organização do programa e da terminologia. Para a harmonização, a combinação da Semântica de *Frames* com a Terminologia baseada em *Frames* (*Frame-based Terminology*) foi fundamental para compreendermos melhor os termos e seu comportamento no contexto do ERP. A natureza experiencial e enciclopédica dos *frames* favoreceu a tarefa terminológica, uma vez que, dada a característica de amplitude e de riqueza dos significados das palavras, foi possível considerar os aspectos contextuais, situacionais, cognitivos e comunicativos dos termos, examinando não apenas a definição terminológica, mas tudo o que está atrelado à terminologia. Desse modo, observamos que a Semântica de *Frames* é capaz de não só sistematizar a base de dados, mas também potencializar a tarefa de harmonização de termos e conceitos.

No que cerne à etapa aplicada desta tese, buscamos desenvolver um *Knowledge Graph* (KG) com *frames* semânticos do ERP, explorando estratégias e técnicas de *prompt* para trabalhar a harmonização de termos e conceitos utilizando o ChatGPT. Com o aporte teórico voltado à abordagem semântico-terminológica para harmonização e com a aproximação entre a representação do conhecimento e os *frames* semânticos, elaboramos um KG do ERP e a descrição linguística da arquitetura a partir da estrutura conceptual. Para a automatização da harmonização em um *Large Language Model* (ChatGPT), criamos diferentes estratégias de *prompt* para automatizarmos a harmonização de termos e conceitos. Inicialmente, instruímos o ChatGPT a descrever a tarefa de harmonização para que pudéssemos compreender se esse LLM sabia em que consistia a atividade e qual metodologia ele seguia. Posteriormente, a cada inserção de novas informações na harmonização, desde as definições terminológicas até a verbalização do *Knowledge Graph* com *frames*, observamos que o modelo aprimorava os resultados. Com isso, percebemos que a automatização da nossa abordagem semântico-terminológica para a harmonização de termos e conceitos foi possível tanto pela base teórica desenvolvida, que dá suporte à tarefa com análises e dados mais completos, como pela arquitetura elaborada neste trabalho (*Knowledge Graph* com *frames* semânticos).

Para atingirmos tais objetivos, focamos na interface linguístico-computacional, a fim de solucionarmos impasses de ordem semântico-terminológico atinentes ao desenvolvimento de ferramentas de gestão de empresas. No que cerne à Linguística, ocupamo-nos, inicialmente, dos compromissos da Linguística Cognitiva e da Semântica Cognitiva para, então, tratarmos da Semântica de *Frames*. Em nosso trabalho, a Semântica de *Frames* é uma das teorias basilares, visto que partimos do princípio de que tal abordagem pode servir como princípio organizador de uma base de dados ERP. Essas discussões serviram para compreendermos e refletirmos como a teoria fillmoreana pode organizar e estruturar uma base de dados de termos no contexto empresarial, o ERP e demais áreas que estão integradas nesse *software*. Além disso, a natureza enciclopédica e a característica de perspectivação dos *frames* contribuem para a análise terminológica associada ao *software*, auxiliando na elaboração de uma representação do conhecimento e na descrição linguística de tal arquitetura.

Em um segundo momento, tratamos dos estudos da Terminologia, traçando um histórico da área, em especial discorrendo sobre a Teoria Geral da Terminologia (TGT) e a sua relação com a normalização e padronização dos termos. Essa aproximação entre

a TGT e a normalização terminológica é fundamental para compreendermos melhor como a harmonização de termos e conceitos foi desenvolvida. Posteriormente, aprofundamo-nos na *Frame-based Terminology* (FBT) para estudarmos e analisarmos a terminologia do *software*. Baseando-se nos pressupostos fillmoreanos, a FBT busca compreender os termos a partir da visão enciclopédica e experiencial. Nesse sentido, aproximam-se os aspectos contextuais e o conhecimento prévio para se trabalhar com a terminologia. Por fim, abordamos a harmonização de termos e conceitos (ISO 860:2007), apresentando, primeiramente, as definições e os métodos criados pela ISO. Em seguida, apresentamos a nossa proposta de harmonização de termos e conceitos de base semântico-terminológica. Tais discussões auxiliaram no entendimento de como a linha teórica de Faber, em adição à Semântica de *Frames* de Charles Fillmore, pode contribuir para utilizar os *frames* como princípio organizador do ERP, pensando no contexto especializado. Ao nos debruçarmos na proposta de harmonização de termos e conceitos da ISO, percebemos que os *frames* contribuíram positivamente para harmonizar a terminologia, especialmente por apresentarem o contexto e a visão enciclopédica para organizarmos a estrutura e os relacionamentos dos termos e conceitos.

Para a frente computacional, focamos nos tópicos sobre Inteligência Artificial (IA), Processamento de Linguagem Natural (PLN) e Representação do Conhecimento. Inicialmente, apresentamos as relações entre a IA, a Filosofia, a Matemática, a Psicologia e a Linguística para, depois, traçarmos o histórico a Inteligência Artificial. Em seguida, tratamos do campo de PLN, subárea da Ciência da Computação, da Inteligência Artificial e da Linguística, visto que nosso trabalho se insere na interface linguístico-computacional. Além disso, sendo nosso objetivo automatizar a harmonização em um *Large Language Model* e utilizar uma representação do conhecimento como estratégia de *prompt* para a tarefa de harmonizar termos e conceitos, discorremos sobre os modelos de linguagem, em especial os LLMs. Por fim, exploramos os conceitos de ontologia, *Knowledge Graph* e *frames* para arquitetarmos a nossa representação do conhecimento do ERP.

Considerando os estudos e as aplicações deste trabalho, os achados desta tese têm diversas implicações tanto práticas quanto teóricas que merecem destaque. Teoricamente, do ponto de vista linguístico, podemos assegurar que os *frames* semânticos servem como princípio organizar de uma base de dados ERP e que uma base semântico-terminológica para a harmonização potencializa a tarefa. Em nossa pesquisa,

observamos que a estrutura conceptual facilitou o mapeamento do *software* e das áreas que o compõem. Ao construirmos a nossa representação do conhecimento do ERP por meio de um *Knowledge Graph* com *frames*, notamos que o conteúdo proveniente das descrições linguísticas dos frames contribuiu para a organização do *software*.

Além disso, no que se refere à Terminologia, esta pesquisa contribui para a literatura existente ao fornecer evidências significativas sobre a relevância de se adicionarem camadas contextuais e cognitivas no procedimento de harmonização de termos e conceitos. Enquanto a ISO não considera o aspecto contextual, a visão enciclopédica e o conhecimento prévio para a análise da terminologia e para a aplicação da harmonização de termos e conceitos, nosso estudo evidencia que trabalhar com essa camada de informações advindas dos *frames* possibilita que se harmonize um conjunto de termos de forma mais completa. Em nossos experimentos, percebemos que apenas as definições terminológicas não seriam suficientes para harmonizar os termos. Desse modo, ao coletarmos mais dados para a análise da terminologia, observamos que a seleção do termo principal e das variantes não era uma tarefa tão simples e direta como se supõe nos procedimentos da ISO. Nessa tarefa, exige-se um conhecimento profundo tanto dos termos do campo a ser trabalhado como da estruturação do domínio para que seja possível elaborar uma harmonização.

Observamos que a tarefa de harmonizar se tornou mais complexa, pois não trabalhamos com os termos de um único domínio. Para a ISO, só é possível harmonizar termos de uma mesma área. Contudo, para harmonizar os dados do ERP, por exemplo, é preciso lidar com diferentes campos, como Compra e Vendas, Recursos Humanos, Finanças, Logística etc. Nesse sentido, as informações provenientes da estrutura conceptual são fundamentais para organizar os domínios do ERP e os termos de cada setor. Com isso, podemos afirmar que a nossa abordagem semântico-terminológica para a harmonização de termos e conceitos torna a proposta da ISO mais robusta.

No que cerne às contribuições teóricas computacionais, esta tese aponta para a importância da elaboração de uma representação do conhecimento para tornar o conteúdo linguístico legível por máquina e da utilização de tal arquitetura nas estratégias e as técnicas de *prompt* em um LLM. Em nossa proposta de harmonização, o desenvolvimento de um *Knowledge Graph* com *frames* semânticos foi fundamental tanto para compreendermos o conjunto de termos no ERP, considerando tanto a estrutura do *software* e dos domínios que o compõem como o conteúdo linguístico atrelado à terminologia a ser analisada e trabalhada. Ao visualizarmos as relações entre



os domínios e os termos, a aplicação da harmonização se tornou um pouco mais clara e acessível. Além disso, essa arquitetura também facilitou a implementação da harmonização em uma base de dados empresarial. Se considerarmos a necessidade de empresas e instituições em compartilharem seus dados entre as diferentes áreas de seu negócio (*Data Fabric*), uma representação desse porte soluciona tal demanda.

Outro ponto se refere ao uso de um *Large Language Model* (LLM). Em estudos prévios, observou-se que os parâmetros de um LLM são fundamentais para obter respostas mais precisas e acuradas. Nesse sentido, estabelecer a temperatura do modelo de linguagem implica diretamente na forma como o LLM irá responder à instrução. Corroborando com pesquisas anteriores, este trabalho amplia o entendimento acerca do uso de *Knowledge Graphs* (KG) em *Large Language Model*. Considerando os resultados obtidos e as análises feitas neste trabalho, podemos observar que a nossa representação do conhecimento (*Knowledge Graph* com *frames* semânticos) melhorou o desempenho do modelo de linguagem, visto que ele apresentou uma harmonização semelhante à nossa mesmo em temperatura alta (temperatura 1). Sabe-se que a temperatura alta pode fazer com que o LLM alucine, gerando dados divergentes à instrução dada. Nossos resultados evidenciaram que, indo ao encontro do que já foi estudado, um *Large Language Model* desenvolve melhor uma tarefa quando se insere mais informações extras, além de uma representação no *prompt*. À vista disso, observamos que o modelo em temperatura mais alta conseguiu relacionar o contexto proveniente da nossa representação do conhecimento com as definições terminológicas e a sua característica criativa na resposta.

É importante salientar que, neste trabalho, não performamos os *prompts* criados para a tarefa de harmonização de termos e conceitos mais de uma vez. Atualmente, sabe-se da necessidade de reproduzir o mesmo *prompt* diversas vezes para avaliar se o comportamento e a resposta gerada pelo modelo de linguagem se mantêm. Essa avaliação se refere à técnica de *Self-Consistency* ou *Self Checking* (Turpin *et al.*, 2023; Fluri; Paleka; Tramèr, 2023; Bartsch; Jorgensen; Rosati, 2023; Wang *et al.*, 2024; Li, Li, Zhang, 2024), na qual a mesma instrução é utilizada mais de uma vez para avaliar se a resposta fornecida pelo modelo é consistente ou se ela varia. O uso de tal técnica é recorrente, haja vista que, nem sempre, os LLMs são fiéis às respostas que fornecem, podendo, em alguns casos, mudar seu raciocínio em mais de uma rodada do mesmo *prompt*. De acordo com Bartsch, Jorgensen e Rosati (2023), os modelos de linguagem podem aprender a expressar inconsistências de pensamentos e crenças devido aos

distintos grupos de pessoas que os geram e os treinam, sobretudo em simulações que exigem do LLM uma cadeia de pensamento ou explicações sobre o seu raciocínio. Para suprir essa problemática, busca-se rodar o mesmo *prompt* não só para testar a eficiência da instrução, mas também para avaliar se os resultados são consistentes em todas as rodadas.

Em termos práticos, é fulcral destacarmos dois pontos. Primeiramente, este trabalho ressalta a importância da entrada de linguistas em trabalhos tecnológicos desenvolvidos no âmbito empresarial. Geralmente, esse tipo de estudo e de aplicação estava restrito à Computação, posto que a busca por soluções e por melhorias em *softwares* eram demandas computacionais. Contudo, percebe-se que, hoje em dia, as empresas de tecnologia buscam profissionais que possam lidar e resolver impasses de ordem semântico-terminológica. Para a solução desse tipo de demanda, os linguistas oferecem uma visão crítica e analítica dos aspectos da linguagem, proporcionando estudos e modelos semânticos, pragmáticos, fonológicos, sintáticos, discursivos etc., que, muitas vezes, não são tão evidenciados em trabalhos computacionais. Nesse sentido, torna-se fundamental aproximar as áreas da Linguística Aplicada e da Computação Aplicada para investigar tais demandas e implementar soluções em *softwares* de gestão de negócios.

Além dessa entrada na área computacional, salientamos a relevância do trabalho conjunto entre linguistas e especialistas da computação para a automatização de tarefas voltadas ao Processamento de Linguagem Natural. Atualmente, percebe-se que, em trabalhos com modelos de linguagem, não é apenas a definição de estratégias e técnicas de *prompt* que melhora a performance de um LLM, mas também a forma como estruturamos a instrução a ser solicitada ao modelo (Liu *et al.*, 2023; John, 2023; Adelson, 2023). Nesse sentido, os profissionais da Linguística podem contribuir em trabalho com LLMs, visto que, para a elaboração de instruções no *prompt*, geralmente, é preciso saber estruturar bem as sentenças, concatenando bem as ideias e deixando a solicitação clara e concisa para se obter os melhores resultados.

Em consonância a isso, automatizar uma tarefa terminológica, como a harmonização de termos e conceitos, exige um olhar da Linguística para que consiga implementar tal abordagem de forma precisa, fazendo os ajustes necessários, a fim de treinar o LLM para realizar a atividade plenamente. Em nossos experimentos, percebemos que o ChatGPT já conhecia a proposta da ISO e se baseia nos pressupostos da Organização para realizar a tarefa. Isso pode ser visto por meio das estratégias de

*prompt* que desenvolvemos para harmonizar o conjunto de termos. Após compreendermos a forma que o LLM harmonizava por si só, elaboramos estratégias para que a nossa proposta de base semântico-terminológica para a harmonização fosse automatizada. Observamos que, a cada nova inserção de informações, o modelo desenvolvia uma resposta mais próxima à nossa harmonização feita manualmente. Quando verbalizamos o nosso *Knowledge Graph* com *frames* semânticos e inserimos a representação junto às definições e aos exemplos, notamos que o LLM apresentou uma harmonização similar a nossa. Com isso, em nível prático, apresentamos uma sequência de possíveis instruções que auxiliam na automatização de tarefas terminológicas, as quais podem servir de base para outras aplicações no âmbito da Terminologia e da Computação.

Quanto às pesquisas futuras e aos desdobramentos relacionados a este trabalho, sugerimos explorar a aplicação manual da abordagem semântico-terminológica para a harmonização de termos e conceitos em contextos diferentes. No que se refere ao uso da nossa proposta de harmonização manual, seria válido verificar se a abordagem necessita de ajustes quando aplicada em dois ou mais domínios para além do campo empresarial. Em nosso trabalho, percebemos que a implementação da harmonização no contexto do ERP, ainda que o *software* seja composto diversos setores, apresenta associações entre as áreas dos pontos de vista terminológico e conceitual. Para que seja possível validar a eficiência da nossa abordagem, sugere-se aplicar a metodologia semântico-terminológica para a harmonização em dois ou mais domínios distintos, mas que, apesar de apresentarem diferenças composicionais, indiquem uma possível relação terminológica entre eles. Um exemplo dessa implementação em domínios distintos, mas com uma certa relação de termos e conceitos, poderia ser feita nos campos da Saúde e da Administração. Se considerarmos que, em um hospital, denominamos uma pessoa internada como ‘paciente’ e este mesmo indivíduo, no setor administrativo, é chamado de ‘cliente’, poderíamos investigar quais outros termos nessas áreas apresentam esse tipo de associação e refletir se há a possibilidade de realizar uma harmonização terminológica.

Ademais, investigações adicionais poderiam examinar a implementação da harmonização de termos e conceitos de base semântico-terminológica automaticamente, em um *Large Language Model*, no contexto tradutório. Como vimos em um pequeno estudo desenvolvido neste trabalho acerca da comparação da harmonização em inglês e

em português<sup>67</sup>, observamos que os resultados das harmonizações não indicaram o mesmo termo principal e suas variantes. Acreditamos que essa diferença entre os resultados da tarefa não seja aleatória, visto que o modelo passa a considerar os aspectos sociais, cognitivos, culturais, históricos, legais e institucionais atrelados às línguas. Conforme afirma Halverson (2013), a Semântica de *Frames* contribui para os estudos tradutórios ao apresentar as influências sociais, históricas e contextuais relacionadas ao conhecimento prévio do tradutor e aos processos cognitivos. Nesse sentido, durante o processo tradutório, o conhecimento enciclopédico auxilia o tradutor nas tomadas de decisão e nas escolhas dos equivalentes na língua de chegada. À vista disso, seria oportuno investigar mais sobre como trabalhar nossa proposta semântico-terminológica para a harmonização de termos e conceitos na tradução.

Além disso, outra sugestão de trabalhos futuros condiz com o uso de *Large Language Models* treinados em língua portuguesa, como o *Maritalk* (Maritaca AI, 2023), para a tarefa de harmonização de termos e conceitos em português. Em nosso trabalho, desenvolvemos e aplicamos a abordagem semântico-terminológica para a harmonização de termos em língua portuguesa utilizando o ChatGPT. Após a finalização desta tese, encontramos o modelo de linguagem *Maritalk*. Iniciamos alguns experimentos de harmonização dos termos em português e obtivemos resultados próximos às harmonizações realizadas manualmente. Esse tipo de trabalho com LLMs treinados em língua portuguesa é uma possibilidade de pesquisa frutífera não só para a avaliação da eficácia da nossa abordagem metodológica, mas também para apoiar as pesquisas em Inteligência Artificial e Linguística Computacional de LLMs especializados no português, desenvolvidos tanto no âmbito de empresas de tecnologia que focam nessa área como no campo acadêmico.

Esta tese, portanto, não só atinge seus objetivos e respalda as premissas elencadas, como também abre novas perspectivas de investigação e aplicação. Pelo viés da Linguística, salientamos a importância da Semântica de *Frames* no sentido de organizar a terminologia de uma base de dados ERP e mapear os domínios do *software*, a fim de representar o conhecimento da área, contribuindo para a elaboração de uma metodologia semântico-terminológica para a harmonização de termos e conceitos. Do ponto de vista da Computação, evidenciamos a relevância de se automatizar a proposta metodológica desenvolvida nesta pesquisa ao utilizar um *Large Language Model* como

---

<sup>67</sup> Um estudo comparativo entre as harmonizações em língua inglesa e em língua portuguesa foi apresentada no capítulo de Metodologia.

recurso computacional para implementação da abordagem. Além disso, apresentamos uma forma de arquitetar um *Knowledge Graph* com *frames semânticos*, demonstrando que esse tipo de representação do conhecimento permite que se estudem e se analisem termos e conceitos. Por fim, contribuímos com a elaboração de diferentes estratégias e técnicas de *prompt* para tarefas de cunho terminológico, mostrando que é possível desenvolver trabalhos de Terminologia na atual conjuntura digital e tecnológica.

## Referências bibliográficas

A. J. Cañas, G. Hill, R. Carff, N. Suri, J. Lott, G. Gómez, T. Eskridge, M. Arroyo, R. Carvajal, in **Concept Maps: Theory, Methodology, Technology**. Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain (September 14-17, 2004), Editorial Universidad Pública de Navarra.

ADELSON, Mark. **The ChatGPT GoldRush Profiting from the AI Revolution Online Prompt Engineering Mastery with ChatGPT**. United Kingdom: Fiola Publishing, 2023.

ADHIKARI, Kamala *et al.* Data harmonization and data pooling from cohort studies: a practical approach for data management. **International journal of population data science**, Wales, v. 6, n. 1, 2021.

AGARWAL, Apoorv; BAUER, Daniel; RAMBOW, Owen. Using frame semantics in natural language processing. Proceedings [...]. Association for Computational Linguistics: Baltimore, 2014. p. 30-33. Disponível em: <https://aclanthology.org/W14-3008.pdf> Acessado em: 17 out 2024.

ALLAN, Keith. **Concise encyclopedia of semantics**. United Kingdom: Elsevier, 2010.

ALLEN, Bradley P.; STORK, Lise; GROTH, Paul. Knowledge Engineering using Large Language Models. [*S.l.*] **arXiv preprint arXiv:2310.00637**, 2023.

ALLMANG, Dean; HENDLER, James; GANDON, Fabien. **Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL**. New York: ACM Books: Association for Computing Machinery, 2019.

ALTO, Valentina. **Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4**. United Kingdom: Packt Publishing Ltd, 2023.

ARP, Robert; SMITH, Barry; SPEAR, Andrew D. **Building ontologies with basic formal ontology**. Massachusetts: MIT Press, 2015.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). **ABNT NBR 13789: Terminologia – Princípios e métodos – Elaboração e apresentação de normas de terminologia**. Rio de Janeiro: ABNT, 1987.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). **ABNT NBR 13790: Terminologia – Princípios e métodos – Harmonização de conceitos e termos**. Rio de Janeiro: ABNT, 1997.

AXELSSON, Agnes; SKANTZE, Gabriel. Using large language models for zero-shot natural language generation from knowledge graphs. [*S.l.*] **arXiv preprint arXiv:2307.07312**, 2023.

AZEVEDO, Diego Napoleão Viana. Teorias da terminologia: descrição, prescrição e explicação. **Revista Cadernos de Tradução**, Porto Alegre, v. 39, n. 3, p. 507-558, 2019.

BARCELONA, Antonio; VALENZUELA, Javier. **An overview of cognitive linguistics**. In.: BRDAR, Mario. *Cognitive Linguistics: Convergence and Expansion*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2011.

BARRASA, Jesús; NATARAJAN, Maya; WEBBER, Jim. **Building Knowledge Graphs: A Practitioner's Guide**. California: O'Reilly, 2022.

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

BARTSCH, Henning *et al.* Self-Consistency of Large Language Models under Ambiguity. [S.l.] **arXiv preprint arXiv:2310.13439**, 2023.

BARZDINĻ, Guntis *et al.*  $\mu$ -Ontologies: Integration of Frame Semantics and Ontological Semantics. Barcelona. **Proceedings [...]**. Barcelona: Sèrie Activitats, 2008, p. 277-283.

BĀRZDIŅŠ, Guntis *et al.* Multidimensional ontologies: integration of Frame semantics and ontological semantics. In: *Proceedings of the 13th EURALEX International Congress*. 2008, p. 277-283.

BATTISTELLA, E. *et al.* Using Protégé to build a Molecular Network Ontology. Madrid. **Proceedings [...]**. Madrid, 8<sup>th</sup> International Protégé Conference, 2005, p. 122.

BELLOMARINI, Luigi *et al.* Knowledge graphs and enterprise AI: the promise of an enabling technology. Macau. **Proceedings [...]**. Macau: IEEE, 2019, p. 26-37.

BERBER SARDINHA, T. Linguística de Corpus: Histórico e Problemática. **Revista DELTA**, São Paulo, vol 16, n.2, 2000, 323-367.

BERGMAN, Erik *et al.* A natural language processing approach towards harmonisation of European medicinal product information. **Plos One**, San Francisco, v. 17, n. 10, p. e0275386, 2022.

BERGOVEC, Marina; RUNJAIĆ, Siniša. Harmonization of Multiple Entries in the Terminology Database Struna (Croatian Special Field Terminology). Madrid. **Proceedings [...]**. Madrid, TKE 2012, 2012, p. 231-241.

BHARGAVA, Cherry; SHARMA, Pradeep Kumar (Ed.). **Artificial intelligence: fundamentals and applications**. Estados Unidos: CRC Press, 2021.

BIEL, Łucja. **Variation of legal terms in monolingual and multilingual contexts**. In.: BIEL, Łucja. *Handbook Of Terminology*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2023.

BLUMAUER, Andreas; NAGY, Helmut. **The knowledge graph cookbook: Recipes that work**. Austria: Edition mono/monochrom, 2020.

BOAS, Hans C. **Computational resources: FrameNet and construction.** In.: DANCYGIER, B. *The Cambridge handbook of cognitive linguistics.* Vancouver: Cambridge, 2017, v. 549, p. 573.

BOLSHAKOV, Igor; GELBUKH, Alexander. **Computational linguistics models, resources, applications.** México: Centro de Investigación en Computación, Instituto Politécnico Nacional, 2004.

BOTSCHEN, Teresa. **Uni-and Multimodal and Structured Representations for Modeling Frame Semantics.** Tese de Doutorado (Doutorado em Ciência da Computação) - Technische Universität Darmstadt, Darmstadt, 2019. Disponível em: <https://tuprints.ulb.tu-darmstadt.de/8484/> Acesso em: 24 ago. 2024.

BOURIGAULT, Didier *et al.* Por uma terminologia textual. **Cadernos de tradução,** Porto Alegre, RS. N. 17 (out./dez. 2004), p. 101-108, 2004.

BOWKER, Lynne. **Terminology management.** In.: ANGELONE, E.; EHRENSBERGER-DOW, G.M. *The Bloomsbury companion to language industry studies,* New York: Bloomsbury Publishing, 2020, p. 261-284.

BOWKER, Lynne. **Terminology.** In: BAKER, M.; SALDANHA, G. *Routledge encyclopedia of translation studies.* Routledge, 2019. p. 579-583.

BRADAR, M.; GRIES, S. **Cognitive Linguistics: Convergence and Expansion.** Amsterdam/Philadelphia: John Publishing, 2011.

BRASOVEANU, Adrian MP *et al.* Framing Few-Shot Knowledge Graph Completion with Large Language Models. Leipzig. **Proceedings [...].** Leipzig: CEUR Workshop Proceedings, 2023.

BRDAR, Mario; FUCHS, Milena Žic; RAFFAELLI, Ida (Ed.). **Cognitive linguistics between universality and variation.** Cambridge: Cambridge Scholars Publishing, 2012.

BREZINA, V., WEILL-TESSIER, P., MCENERY, A. (2020). **#LancsBox v. 5.x.** [software]. Available at: <http://corpora.lancs.ac.uk/lancsbox>

BREZINA, Vaclav. **Statistics in corpus linguistics: A practical guide.** London: Cambridge University Press, 2018.

CABRÉ, M.T.. A terminologia hoje: concepções, tendências e aplicações. Tradução Susana Kerschner. **Revista Cadernos de Tradução, A Terminologia em foco,** Porto Alegre, n 17, out.-dez., 2004, 9-30.

CABRÉ, M.T.. **Terminology: Theory, methods and applications.** Philadelphia: John Benjamins Publishing Company, 1998.

CABRÉ, María Teresa; CASAFONT, Mercè Lorente. Panorama teòric de la terminologia actual. **Revista Terminalia,** Barcelona, n. 24, p. 57-63, 2021.



CABRÉ, Maria Teresa. **La terminología**: teoria, metodologia, aplicaciones. Barcelona: Editorial Antárdita/Wempúries, 1993.

CAI, Zixing *et al.* **Artificial Intelligence**: From Beginning to Date. China: Tsinghua University Press, 2021.

CAMPESATO, Oswald. **Artificial intelligence, machine learning, and deep learning**. Estados Unidos: Mercury Learning and Information, 2020.

CAMPESATO, Oswald. **Transformer, BERT, and GPT**: Including ChatGPT and Prompt Engineering. Estados Unidos: Mercury Learning and Information, 2023.

CAÑAS, A.J.; HILL, G.; CARFF, R.; LOTT, J.; GÓMEZ, G.; ESKRIDGE, T.; ARROYO, M.; CARVAJAL, R. **CmapTools: A knowledge modelling and sharing environment**. In: Concept Maps: Theory, methodology, technology. Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain (set. 14-17, 2004), Editorial Universidad Pública de Navarra, 2004.

CASELI, H.M.; NUNES, M.G.V.; PAGANO, A. **O que é PLN?** In.: CASELI, H.M.; NUNES, M.G.V. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português – 2ª. Edição*. São Carlos: BPLN, 2024. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-introducao/cap-introducao/cap-introducao.html>

CASTELLVÍ, M. Teresa Cabré. Theories of terminology: Their description, prescription and explanation. Terminology. **International Journal of Theoretical and Applied Issues in Specialized Communication**, Amsterdam/Philadelphia, v. 9, n. 2, p. 163-199, 2003.

CHAKRABORTY, Utpal; ROY, Soumyadeep; KUMAR, Sumit. **Rise of Generative AI and ChatGPT**: Understand how Generative AI and ChatGPT are transforming and reshaping the business world (English Edition). Noida: BPB Publications, 2023.

CHARNIAK, E. **Introduction to Deep Learning**. London: MIT Press, 2018.

CHEIN, Michel; MUGNIER, Marie-Laure. **Graph-based knowledge representation**: computational foundations of conceptual graphs. Berlin: Springer Science & Business Media, 2008.

CHEN, Zixi *et al.* What makes a good explanation?: A harmonized view of properties of explanations. [*S.l.*] **arXiv preprint arXiv:2211.05667**, 2022.

CHIAVEGATTO, V.C.. Introdução à Linguística Cognitiva. **Revista Matraca**, Rio de Janeiro, v.16, n.24, jan./jun., 2009.

CHINESE ACADEMY of ENGINEERING. **The Development of Natural Language Processing**: Research on the Development of Electronic Information Engineering Technology in China. Singapore: Springer Singapore, 2021.

CHISHMAN, Rove Luiza de Oliveira *et al.* **Dicionário Olímpico**. São Leopoldo: Unisinos, 2016. Disponível em: <https://www.dicionarioolimpico.com.br/> Acesso em: 24 ago. 2024

CHISHMAN, Rove Luiza de Oliveira *et al.* **Dicionário Paralímpico**. São Leopoldo: Unisinos, 2021. Disponível em: <https://www.dicionarioparalimpico.com.br/>

CHISHMAN, Rove Luiza de Oliveira *et al.* **Field - dicionário de expressões do futebol**. São Leopoldo: Unisinos, 2014. Disponível em: <http://dicionariofield.com.br/>

CHISHMAN, R.L. *et al.* Cenário Gol. In.: CHISHMAN, Rove Luiza de Oliveira *et al.* **Field - dicionário de expressões do futebol**. São Leopoldo: Unisinos, 2014. Disponível em: <http://dicionariofield.com.br/>

CHISHMAN, R.L. *et al.* **Gol de Letra**. In.: CHISHMAN, Rove Luiza de Oliveira *et al.* **Field - dicionário de expressões do futebol**. São Leopoldo: Unisinos, 2014. Disponível em: <http://dicionariofield.com.br/>

CHISHMAN, R.L. *et al.* **Voleibol**. In.: CHISHMAN, Rove Luiza de Oliveira *et al.* **Dicionário Olímpico**. São Leopoldo: Unisinos, 2016. Disponível em: <https://www.dicionarioolimpico.com.br/>

CHISHMAN, R.L. *et al.* **Cenário Ataque**. In.: CHISHMAN, Rove Luiza de Oliveira *et al.* **Dicionário Olímpico**. São Leopoldo: Unisinos, 2016. Disponível em: <https://www.dicionarioolimpico.com.br/>

CHISHMAN, R.L. *et al.* **Bomba**. In.: CHISHMAN, Rove Luiza de Oliveira *et al.* **Dicionário Olímpico**. São Leopoldo: Unisinos, 2016. Disponível em: <https://www.dicionarioolimpico.com.br/>

CHISHMAN, R.L. *et al.* **Natação**. In.: CHISHMAN, Rove Luiza de Oliveira *et al.* **Dicionário Paralímpico**. São Leopoldo: Unisinos, 2021. Disponível em: <https://www.dicionarioparalimpico.com.br/>

CHISHMAN, R.L. *et al.* **Cenário Equipamentos**. In.: CHISHMAN, Rove Luiza de Oliveira *et al.* **Dicionário Paralímpico**. São Leopoldo: Unisinos, 2021. Disponível em: <https://www.dicionarioparalimpico.com.br/>

CHISHMAN, R.L. *et al.* **Tapper**. In.: CHISHMAN, Rove Luiza de Oliveira *et al.* **Dicionário Paralímpico**. São Leopoldo: Unisinos, 2021. Disponível em: <https://www.dicionarioparalimpico.com.br/>

CHOMSKY, N. **Syntactic Structures**. The Hague: Mouton, 1957.

CHOMSKY, Noam. Logical structure in language. **Journal of the American Society for Information Science**, New York, v. 8, n. 4, p. 284, 1957.

CHOUDHARY, Nurendra; REDDY, Chandan K. Complex logical reasoning over knowledge graphs using large language models. [S.l.] **arXiv preprint arXiv:2305.01157**, 2023.

CIENKI, A. **Frames, Idealized Cognitive Models, and Domains**. In: GEERAERTS, D.; CUYCKENS, H. *The Oxford Handbook of Cognitive Linguistics*. New York: Oxford University Press, 2007.

CLARK, A.; FOX, C.; LAPPIN, S. **The Handbook of Computational Linguistics and Natural Language Processing**. Oxford: John Wiley & Sons, 2010.

CmapTools: A Knowledge Modeling and Sharing Environment, A. J. Cañas, G. Hill, R. Carff, N. Suri, J. Lott, G. Gómez, T. Eskridge, M. Arroyo, R. Carvajal, in *Concept Maps: Theory, Methodology, Technology, Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain (September 14-17, 2004)*, Editorial Universidad Pública de Navarra.

CORCOGLIONITI, Francesco; ROSPOCHER, Marco; APROSIO, Alessio Palmero. Frame-based ontology population with PIKES. **IEEE Transactions on Knowledge and Data Engineering**, United States, v. 28, n. 12, p. 3261-3275, 2016.

CORREA, L.M.S. **Língua e Cognição: Antes e depois da revolução cognitiva**. In: GUIMARÃES, E (ed.). *Introdução às ciências da linguagem: linguagem história e conhecimento*. Campinas: Editoras Pontes, 2006.

COSTA, Gustavo de Assis; DE OLIVEIRA, José Maria Parente. Linguistic frames as support for entity alignment in knowledge graphs. Linz. **Proceedings [...]**. Linz: Association for Computing Machinery, 2018, p. 226-229.

COZMAN, Fabio G.; PLONSKI, Guilherme Ary; NERI, Hugo. **Inteligência artificial: avanços e tendências**. Universidade de São Paulo. Instituto de Estudos Avançados, 2021. Disponível em: [www.livrosabertos.abcd.usp.br/portaldelivrosUSP/catalog/book/650](http://www.livrosabertos.abcd.usp.br/portaldelivrosUSP/catalog/book/650). Acesso em 16 outubro. 2024.

CRAM, Damien; DAILLE, Béatrice. Terminology extraction with term variant detection. Berlin. **Proceedings [...]**. Berlin: Association for Computational Linguistics, 2016.

CROFT, W.; CRUSE, D. A. **Cognitive Linguistics**. Cambridge: Cambridge University Press, 2004.

CURTI-CONTESSOTO, Beatriz; COSTA, Lucimara Alves. Terminologia e terminólogos: teorias, aplicabilidades e mercado de trabalho. **Digital Internacional de Lexicología, Lexicografía y Terminología**, Córdoba, n. 4, 2021.

CZULO, Oliver; ZIEM, Alexander; TORRENT, Tiago Timponi. Beyond lexical semantics: notes on pragmatic frames. Marseille. **Proceedings [...]**. Marseille: European Language Resources Association, 2020.

DĄBROWSKA, Ewa; DIVJAK, Dagmar (Ed.). **Cognitive Linguistics-Foundations of Language**. Berlin: Walter de Gruyter GmbH & Co KG, 2019.

DAILLE, Béatrice. **Term variation in specialised corpora**. Amsterdam: John Benjamins Publishing Company, 2017.

Data Science Academy. **Deep Learning Book**, 2022. Disponível em: <<https://www.deeplearningbook.com.br/>>. Acesso em: 24 dez. 2023.

DE FERNANDES TEIXEIRA, João. **Inteligência artificial**. São Paulo: Pia Sociedade de São Paulo-Editora Paulus, 2014.

DE FREITAS, Larissa Astrogildo; VIEIRA, Renata. Ontologias e língua portuguesa. Pelotas. **Anais** [...]. Pelotas: EDUCAT, v. 1, n. 2, p. 5, 2008.

DE SCHRYVER, Gilles-Maurice. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. **International Journal of Lexicography**, Oxford, v. 36, n. 4, p. 355-387, 2023.

DELAVIGNE, Valérie; GAUDIN, François. **Founding principles of Socioterminology**. In.: FABER, P.; L'HOMME, M.C. Theoretical Perspectives on Terminology. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2022, p. 177-196.

DENG, L.; LIU, Y. **Deep Learning in Natural Language Processing**. Singapore: Springer, 2018.

DEY, Priyanka; GIRJU, Roxana. Enriching deep learning with frame semantics for empathy classification in medical narrative essays. Abu Dhabi. **Proceedings** [...]. Abu Dhabi: Association for Computational Linguistics, 2022. p. 207-217.

DI FELIPPO, A. **Ontologias Linguísticas aplicadas ao processamento automático das línguas naturais: o caso das redes wordnets**. In: MAGALHÃES, J. S.; TRAVAGLIA, L. C. Múltiplas perspectivas em Linguística. Uberlândia: Edufu, 2008.

DiCoEnviro. **Dictionnaire fondamental de l'environnement**. Disponível em: <http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search.cgi>. Acesso em: 20 abr. 2022.

DIEDERICH, Catherine. **Chapter sixteen: the use of framing to conceptualize specialized terminology**. In: GRYGIEL, M. COGNITIVE approaches to specialist languages. Newcastle: Lady Stephenson Publishing, 2017, p. 351.

DUNN, J. **Natural Language Processing for Corpus Linguistics**. United Kingdom: Cambridge University Press, 2022.

EcoLexicon. **EcoLexicon Terminological knowledge base**. Disponível em: <http://ecolexicon.ugr.es/en/index.htm> Acesso em abr. 2022.

EGBERT, Jesse; LARSSON, Tove; BIBER, Douglas. **Doing linguistics with a corpus: Methodological considerations for the everyday user**. Cambridge: Cambridge University Press, 2020.

EVANS, V.; GREEN, M. **Cognitive Linguistics: an introduction**. Edinburgh: Edinburgh University Press, 2006.

EVANS, Vyvyan. **Cognitive linguistics: A complete guide**. Edinburgh: Edinburgh University Press, 2019.

EVANS, Vyvyan. **The crucible of language: How language and mind create meaning**. Cambridge: Cambridge University Press, 2015.

FABER, P; CABEZAS-GARCÍA, M.. Specialized Knowledge Representation: From Terms to Frames. **Research in Language**, Lodz, 17(2), 197-211, 2019.

FABER, P; LEÓN-ARÁUZ, P.. Transforming glossaries into knowledge resources: frame-based terminology applied to military science. Elazig. **Proceedings [...]**. Elazig: Asos Publisher, 2019, p. 412-438.

FABER, P. **A cognitive linguistics view of terminology and specialized language**. Berlin, Boston: De Gruyter Mouton, 2012.

FABER, P.; JIMÉNEZ HURTADO, C. **Traducción, lenguaje y cognición**. Granada: Comares, 2004.

FABER, P.; L'HOMME, M.C. Lexical semantic approaches to terminology. An introduction. **Terminology: international journal of theoretical and applied issues in specialized communication**, 20(2):143-150. Amsterdam/Philadelphia: John Benjamins, 2014.

FABER, P.; LEÓN ARAÚZ, P. Dinamismo conceptual en las bases de conocimiento terminológico: el caso de EcoLexicon. **Íkala, Revista de lenguaje y cultura**, Colombia, 15(25), p.75-100, 2010.

FABER, P.; LEÓN ARAUZ, P.; PRIETO VELASCO, J.A.; REIMERINK, A. Linking images and words: the description of specialized concepts. Torino. **Proceedings [...]**. Torino: AbeBooks, 2006, p. 751-763.

FABER, P.; LEÓN ARAÚZ, P.; PRIETO VELASCO, J.A.; REIMERINK, A. Linking images and words: the description of specialized concepts. **International Journal of Lexicography**, Oxford, 20(1):39-65. Oxford Univ Press, 2007.

FABER, P.; LEÓN-ARAÚZ, P.; REIMERINK, A.. EcoLexicon: new features and challenges. Slovenia. **Proceedings [...]**. Slovenia: GLOBALEX 2016, 2016, p. 73-80.

FABER, P.; LEÓN-ARÁUZ, P.. **Frame-based terminology Applied to military Science: transforming a glossary into a knowledge resource**. Berlin: Springer, 2019.

FABER, P.; MÁRQUEZ LINARES, C. **The role of imagery in specialised communication**. In.: LEWANDOWSKA-TOMASZCZYK, B.; KWIATKOWSKA, A. *Imagery in language*. Frankfurt: Peter Lang, 2004, P. 585-602.

FABER, P.; MÁRQUEZ LINARES, C.; VEGA EXPÓSITO, M. **Framing Terminology: A process-oriented approach.** Meta: Journal des traducteurs / Meta: Translators' Journal, 50(4). Montreal: Les Presses de l'Université de Montréal, 2005.

FABER, P.; MÁRQUEZ LINARES, C.; VEGA EXPÓSITO, M. **Modelización conceptual en la traducción científica.** In: EMSEL, M.; OTAL, C. Brücken: Festschrift für Gerd Wotjak zum 65. Geburtstag. Frankfurt: Peter Lang, 2007, p. 109-120.

FABER, P.; MONTERO MARTÍNEZ, S.; CASTRO PRIETO, M.R.; SENSO RUIZ, J.; PRIETO VELASCO, J.A.; LEÓN ARAUZ, P.; MÁRQUEZ LINARES, C.; VEGA EXPÓSITO, M. Process-oriented terminology management in the domain of Coastal Engineering. **Terminology**, Amsterdam/Philadelphia, 12(2):189-213, 2006.

FABER, P.. **Frames as a framework for terminology.** In: KOCKART, H.J; STEURS, F. The Handbook of Terminology 1. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015, p. 14.33.

FABER, Pamela; L'HOMME, Marie-Claude (Ed.). **Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge.** Amsterdam/Philadelphia: John Benjamins Publishing Company, 2022.

FABER, Pamela; L'HOMME, Marie-Claude. Lexical semantic approaches to terminology. **Terminology**, Amsterdam/Philadelphia, v. 20, n. 2, p. 143-150, 2014.

FABER, P. *et al.*. **Climate change.** Universidade de Granada: Granada, 2024.

FACELI, Katti *et al.* **Inteligência artificial: uma abordagem de aprendizado de máquina.** Rio de Janeiro: Editora LTC, 2021.

FAULSTICH, E. Planificação Linguística e Problemas de Normalização. **Revista Alfa**, São Paulo, 42 (n. esp): 247-268, 1998.

FELBER, H. International Standardization of Terminology: theoretical and methodological aspects. **International Journal of the sociology Languages**, Berlin, 23, 65-79, 1980.

FELBER, Helmut. **Basic principles and methods for the preparation of terminology standards.** In: Standardization of technical terminology: Principles and practices. ASTM International, 1983.

FELBER, Helmut. Theory of terminology, terminology work and terminology documentation. **Fachsprache. Internationale Zeitschrift für Fachsprachenforschung,-didaktik und Terminologie Wien**, Wein, v. 1, n. 1, p. 20-32, 1979.

FENG, Chao; ZHANG, Xinyu; FEI, Zichu. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. [S.l.] **arXiv preprint arXiv:2309.03118**, 2023.

FENSEL, Dieter *et al.* **Introduction:** what is a knowledge graph?. Knowledge graphs: Methodology, tools and selected use cases. Berlin: Springer, Cham, 2020, p. 1-10.

FERNANDES, Patrick *et al.* The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. [S.l.] **arXiv preprint arXiv:2308.07286**, 2023.

FERNÁNDEZ-SILVA, S.; FREIXA, J.; CABRÉ, M.T. **A Cognitive approach to synonymy in Terminology.** In.: BRDAR, Mario; FUCHS, Milena Žic; RAFFAELLI, Ida (Ed.). Cognitive linguistics between universality and variation. Cambridge: Cambridge Scholars Publishing, 2012.

FILLMORE, C. J. **An alternative to checklist theories of meaning.** FIRST ANNUAL MEETING OF THE BERKELEY LINGUISTICS SOCIETY. In: Proceedings of the first annual meeting of the Berkeley Linguistics Society. Berkeley: Berkeley Linguistics Society, 1975, p.123-131.

FILLMORE, C. J. **Frame semantics and the nature of language.** In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, Vol. 280, 1976, p. 20-32.

FILLMORE, C. J. **Frame semantics.** In: The Linguistic Society of Korea (Eds.). Linguistics in the Morning Calm. Seoul: Hanshin, 1982, p.111-37.

FILLMORE, C. J. **Frames and the semantics of understanding.** Quaderni di Semantica. Vol.6, N.2, 1985, p.222-254.

FILLMORE, C. J. **Scenes-and-frames semantics.** In ZAMPOLLI, A. (Ed.): Linguistic Structures Processing: Fundamental Studies in Computer Science, No. 59, North Holland Publishing, 1977, p.55-88.

FILLMORE, Charles J. **The case for case.** In: BACH, E; HARMS, R.T. Universals in linguistic theory. New York, NY: Holt, Rinehart, and Winston, 1967.

FILLMORE, C. J.; BAKER, C. **A frames approach to semantic analysis.** In: HEINE, B.; NARROG, H. (eds.). The Oxford Handbook of Linguistic Analysis. New York: Oxford University Press, 2009,. p. 313-339.

FINATTO, M.J; KILLIAN, C. **Texto-comentário 7.** In: FINATTO, M.J.. Textos e termos por Lothar Hoffmann. Porto Alegre: Palotti, 2015.

FINATTO, Maria José Bocorny *et al.* **Conceitos básicos da Lingüística de Linguagens Especializadas–texto comentário 2.** Textos e termos por Lothar Hoffmann, um convite para o estudo das linguagens técnicocientíficas. Porto Alegre: Palotti, 2015.

FINATTO, M.J. **Terminografia brasileira no final do século XIX:** contraponto entre domínios emergentes e consolidados. In: KRIEGER, Maria da Graça; MACIEL, Anna Maria Becker. (org.). Temas de terminologia. São Paulo. Porto Alegre: Ed. Universidade/UFRGS/Humanitas/USP, 2001.

FINATTO, Maria José; ZILIO, Leonardo. **Textos e termos para Lothar Hoffmann: um convite para o estudo das linguagens técnico-científicas.** Porto Alegre, Editora Pallotti/FAPERGS, 2015.

FLURI, Lukas; PALEKA, Daniel; TRAMÈR, Florian. Evaluating superhuman models with consistency checks. Toronto. **Proceedings** [...]. Toronto: IEEE, 2024. p. 194-232.

FOO, Jody; MERKEL, Magnus. Computer aided term bank creation and standardization. **Terminology in Everyday Life**, Amsterdam/Philadelphia, v. 13, p. 163, 2010.

FRANA, Philip L.; KLEIN, Michael J. (Ed.). **Encyclopedia of artificial intelligence: the past, present, and future of AI.** USA: Bloomsbury Publishing, 2021.

FREITAS, Artur; BORDINI, Rafael H.; VIEIRA, Renata. Model-driven engineering of multi-agent systems based on ontologies. **Applied Ontology**, Amsterdam, v. 12, n. 2, p. 157-188, 2017.

FRIGINAL, Eric; HARDY, Jack A. (Ed.). **The Routledge handbook of corpus approaches to discourse analysis.** Routledge, 2020.

FU, Yao *et al.* Complexity-based prompting for multi-step reasoning. [*S.l.*] **Proceedings** [...]. [*S.l.*]: The Eleventh International Conference on Learning Representations, 2022.

GANDON, Fabien L. **Ontologies in Computer Science: These New “Software Components” of our Information Systems.** In: Gargouri F. (Eds.), *Ontology Theory, Management and Design: Advanced tools and Models*, Higher Institute of Informatics and Multimedia of Sfax, Tunisia, 2010.

GANGEMI, Aldo *et al.* Framester: A wide coverage linguistic linked data hub. Bologna. **Proceedings** [...]. Bologna: Springer International Publishing, 2016. p. 239-254.

GAUDIN, F. Socioterminologie. **Des problèmes sémantiques aux pratiques institutionnelles.** Rouen, Université de Rouen, 1993.

GAWRON, Jean Mark *et al.* **Frame semantics.** In.: MAIENBORN, C.; VON HEUSINGER, K.; PORTNER, P. Berlin: *Semantics theories*, p. 57-85, 2009.

GEERAERTS, D. **Cognitive Linguistics: Basic Readings.** Berlin: Mouton de Gruyter, 2006.

GEERAERTS, D. **Handbook of Terminology** (Vol. 1). (H. Kockaert, & F. Steurs, Eds.) Amsterdam. 2015.

GEERAERTS, D. **Lexical Semantics.** In: DĄBROWSKA, Ewa; DIVJAK, Dagmar (Ed.). *Cognitive Linguistics-Foundations of Language.* Berlin: Walter de Gruyter GmbH & Co KG, 2019.



GEERAERTS, D.; CUYCKENS, H. **Introducing Cognitive Linguistics**. In: GEERAERTS, D.; CUYCKENS, H. *The Oxford Handbook of Cognitive Linguistics*. New York: Oxford, 2007.

GEERAERTS, D.; CUYCKENS, H. **The Oxford Handbook of Cognitive Linguistics**. New York: Oxford University Press, 2007.

GEERAERTS, Dirk. **Cognitive semantics**. In: WEN, Xu; TAYLOR, John R. (Ed.). *The Routledge handbook of cognitive linguistics*. London; New York, NY: Routledge, 2021, p. 19-29.

GIACOMINI, Laura; SCHÄFER, Johannes. Computational Aspects of Frame-Based Meaning Representation in Terminology. Marseille. **Proceedings** [...]. Marseille, European Language Resources Association, 2020. p. 80-84.

GÖDERT, Winfried; HUBRICH, Jessica; NAGELSCHMIDT, Matthias. **Semantic knowledge representation for information retrieval**. Alemanha: Walter de Gruyter GmbH & Co KG, 2014.

GOODFELLOW, Ian; YOSHUA, Bengio; COURVILLE, Aaron. **Deep learning**. Massachusetts: MIT press, 2016.

GOYAL, T; BHATTACHARYYA, P. Knowledge Graphs and Knowledge Infusion in Language Models. [S.l.]: **arXiv preprint arXiv:2403.01481**, 2023.

GRANADA, Roger *et al.* A comparable corpus based on aligned multilingual ontologies. Republic of Korea. **Proceedings** [...]. Republic of Korea, Association for Computational Linguistics, 2012, p. 25-31.

GRENIÉ, Matthias *et al.* Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. **Methods in Ecology and Evolution**, London, v. 14, n. 1, p. 12-25, 2023.

GROTH, Paul *et al.* Knowledge graphs and their role in the knowledge engineering of the 21st century. **Dagstuhl Reports**, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Wandering, 2023.

GRYGIEL, Marcin (Ed.). **Cognitive approaches to specialist languages**. Cambridge: Cambridge Scholars Publishing, 2017.

GUO, Shaoru, *et al.* Incorporating syntax and frame semantics in neural network for machine reading comprehension. Barcelona. **Proceedings** [...]. Barcelona, International Committee on Computational Linguistics, 2020.

HAGIWARA, M. **Real-World Natural Language Processing**. New York: Manning Publications, 2021.

HALVERSON, S. **Implications of Cognitive Linguistics for Translation Studies**. In: ROJO, A; IBARRETXE-ANTUÑANO, I. *Cognitive Linguistics and Translation*

advances in some theoretical models and applications. Germany: De Gruyter Company, 2013.

HARDER, Peter. **Meaning in mind and society: A functional contribution to the social turn in cognitive linguistics.** Germany: De Gruyter Mouton, 2010.

HARISPE, Sébastien *et al.* **Semantic similarity from natural language and ontology analysis.** Nimes, France: Morgan & Claypool, 2015.

HAUSSER, R. **Foundations of Computational Linguistics: Human-Computer Communication in Natural Language.** Heidelberg: Springer, 2014.

HE, Zhiwei *et al.* Exploring human-like translation strategy with large language models. **Transactions of the Association for Computational Linguistics**, Massachusetts, v. 12, p. 229-246, 2024.

HERSHCOVICH, Daniel; DONATELLI, Lucia. It's the meaning that counts: the state of the art in NLP and semantics. **KI-Künstliche Intelligenz**, Berlin, v. 35, n. 3, p. 255-270, 2021.

HERWEG, Michael. Dynamic Event Types in Frame Semantics: The Representation of Change in FAMEu. **Heidelberg University Papers on Language & Cognition 2.1**, Heidelberg, 2021, p. 15-44.

HJØRLAND, Birger. Terminology. **KO KNOWLEDGE ORGANIZATION**, Baden-Baden, v. 50, n. 2, p. 111-127, 2023.

HOFFMANN, Lothar. Fachsprache/Language of Specific Purposes. **Sociolinguistics/Soziolinguistik**, Berlin, v. 1, p. 232-238, 2004.

HOFFMANN, Lothar. Fachsprachenforschung/Research on Languages for Special Purposes. **Soziolinguistik: Ein Internationales Handbuch Zur Wissenschaft Von Sprache Und Gesellschaft**, Berlin, v. 2, p. 1385-1393, 2005.

HUANG, Chu-Ren (Ed.). **Ontology and the lexicon: A natural language processing perspective.** Cambridge: Cambridge University Press, 2010.

HUANG, Qing *et al.* Let's Chat to Find the APIs: Connecting Human, LLM and Knowledge Graph through AI Chain. Danvers. **Proceedings [...]**. Danvers, Conference Publishing Services, 2023. p. 471-483.

INDURKHYA, M.; DAMERAU, F. **Handbook of Natural Language Processing.** Cambridge: Taylor & Francis group, 2010.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO); INTERNATIONAL LANGUAGE AND TERMINOLOGY COMMISSION. **ISO 704: Terminology work – Principles and methods.** 2<sup>nd</sup> ed. Geneva: ISO, 2000. (Technical report).

JOHN, Ibrahim. **The art of asking ChatGPT for high-quality answers: a complete guide to prompt engineering techniques.** [S.l.]: Nazunda Technologies Limited, 2023.

KEJRIWAL, M.; KNOBLOCK, C.A.; SZEKELY, P. **Knowledge Graphs: Fundamentals, Techniques, and Applications.** London: MIT Press, 2021.

KEJRIWAL, Mayank; KNOBLOCK, Craig; SZEKELY, Pedro. **Knowledge Graphs: Fundamentals, Techniques, and Applications.** Massachusetts: MIT Press, 2021.

KELLEHER, John D. **Deep learning.** Massachusetts: MIT press, 2019.

KENEDY, Eduardo. **Gerativismo.** In: MARTELOTTA, Mário (Org.) **Manual de Lingüística.** São Paulo: Contexto, p. 127-140, 2008.

KILGARRIFF, Adam *et al.* The Sketch Engine: ten years on. **Lexicography Journal of Asialex,** Amsterdam/Philadelphia, v. 1, p. 7-36, 2014.

KILGARRIFF, A.; RYCHLÝ, P; SMRŽ, P.; TUGWELL, D. **Itri-04-08 the Sketch Engine.** Information Technology, 2004.

KISHORE, Rajiv; RAMESH, Ram (Ed.). **Ontologies: a handbook of principles, concepts and applications in information systems.** Alemanha: Springer Science & Business Media, 2007.

KOCH, I.; CUNHA-LIMA, M.L.. **Do cognitivismo ao sociocognitivismo.** In: MUSSALIM, F.; BENTES, A.C. **Introdução à Linguística 3: fundamentos epistemológicos.** 4ed. São Paulo: Cortez, 2009.

KOCKAERT, Hendrik J.; STEURS, Frieda (Ed.). **Handbook of terminology.** Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015.

KOESTER, A. **Building a small specialized corpora.** In: O'KEEFFEE, A.; MCCARTHY, M. **The Routledge Handbook of Corpus Linguistics.** London and New York: Routledge, 2010.

KOEVA, Svetla; DOYCHEV, Emil. **Ontology supported frame classification.** Bulgaria. **Proceedings [...].** Bulgaria: Department of Computational Linguistics, IBL -- BAS , In: 2022, p. 203-213.

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

KRIEGER, M.G. **Terminologia revisitada.** Revista DELTA, vol, 16, n. 2, 209-228, 2000.

KUBLIK, S.; SABOO, S. **GPT-3: Building Innovative NLP Products Using Large Language Models.** California: O'Reilly Media, 2022.

KUMAR, Avadhesh *et al.* (Ed.). **Prediction and Analysis for Knowledge Representation and Machine Learning.** Flórida: CRC Press, 2022.

KUMAR, Ganesh *et al.* Data harmonization for heterogeneous datasets: a systematic literature review. **Journal Applied Sciences**, [*S.l.*], v. 11, n. 17, p. 8275, 2021.

L'HOMME, M.C. **Lexical Semantics for Terminology: an introduction**. Philadelphia: John Benjamins Publishing Company, 2020.

L'HOMME, M.C. **Management of terminology in a machine-translation environment**. *Revista Terminology*, 1:1, 121-135, 1994.

L'HOMME, M.C.; LANEVILLE, M.E. **El diccionario fundamental del medio ambiente**. Montreal: OLST, 2010.

L'HOMME, M.C. *et al.* **Climate and meteorological events**. Montreal: OLST, 2010.

L'HOMME, M.C.; ROBICHAUD, B. Frames and Terminology: Representing predicative terms in the field of the environment. Dublin. **Proceedings** [...]. Dublin, Association for Computational Linguistics and Dublin City University, p. 186-197, 2014.

LACASTA, Javier; NOGUERAS-ISO, Javier; SORIA, Francisco Javier Zarazaga. **Terminological ontologies: design, management and practical applications**. Alemanha: Springer Science & Business Media, 2010.

LANGACKER, R. **Convergence in Cognitive Linguistics**. In: BRADAR, M.; GRIES, S. *Cognitive Linguistics: Convergence and Expansion*. Amsterdam/Philadelphia: John Publishing, 2011.

LAPPIN, Shalom. **Deep learning and linguistic representation**. United Kingdom: Chapman and Hall/CRC, 2021.

LEIPNITZ, L. **Texto-comentário 3**. In: FINATTO, M.J.. *Textos e termos por Lothar Hoffmann*. Porto Alegre: Palotti, 2015.

LEÓN ARAÚZ, P.; REIMERINK, A.; FABER, P. Puertoterm & Marcocosta: A Frame-based Knowledge Base for the Environmental Domain. **Journal of Multicultural Communication**, White Plains, 1:47-70, 2009.

LEÓN ARAÚZ, P.; REIMERINK, A.; FABER, P. PUERTOTERM and MARCOCOSTA: a frame-based knowledge base for the environmental domain. Shanghai. **Proceedings** [...]. Shanghai, Languages Press, 2008.

LI, Taiji; LI, Zhi; ZHANG, Yin. Improving Faithfulness of Large Language Models in Summarization via Sliding Generation and Self-Consistency. Torino. **Proceedings** [...]. Torino, ELRA and ICCL, 2024. p. 8804-8817.

LI, J. *et al.* Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. **Transactions of the Association for Computational Linguistics**, Massachusetts, 12, 576-592, 2023.

- LI, Zekun *et al.* Guiding large language models via directional stimulus prompting. **Advances in Neural Information Processing Systems**, San Diego, v. 36, 2024.
- LIMA, A.D.F. **Variantes Terminológicas: necessidade de harmonização nos regulamentos internos do Ministério da Ciência e Tecnologia**. 2016. Dissertação (Mestrado em Terminologia e Gestão da Informação de Especialidade) – Programa de Pós-Graduação em Linguística, Universidade Nova de Lisboa, Lisboa, 2016.
- LIMA, I.; PINHEIRO, C.A.M.; SANTOS, F.A.O. **Inteligência Artificial**. Rio de Janeiro: Elsevier Editora, 2014.
- LIN, Asiyah Yu *et al.* A community effort for COVID-19 ontology harmonization. Slovenia. **Proceedings** [...]. Slovenia, NIH Public Access, 2022. p. 122.
- LIPPOLIS, Anna Sofia *et al.* Enhancing Entity Alignment Between Wikidata and ArtGraph using LLMs. Athens. **Proceedings** [...]. Athens, SWODCH, 2023.
- LISBOA, J.V.R. **Proposta de harmonização da terminologia designativa de área e subáreas do português como língua não materna baseada em corpus**. 2021. Dissertação (Mestrado em Estudos Linguísticos) – Programa de Pós-Graduação em Estudos Linguísticos do Instituto de Letras e Linguística, Universidade Federal de Uberlândia, Uberlândia, 2021.
- LIU, Jiacheng *et al.* Generated knowledge prompting for commonsense reasoning. [S.l.] **arXiv preprint arXiv:2110.08387**, 2021.
- LIU, Z.; LIN, Y.; SUN, M. **Representation Learning for Natural Language Processing**. Singapore: Springer, 2020.
- LONG, Jieyi. Large language model guided tree-of-thought. [S.l.] **arXiv preprint arXiv:2305.08291**, 2023.
- LOPES, Lucelene *et al.* Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. **RECHIS: electronic journal of communication & innovation in health**. Rio de Janeiro, RJ. Vol. 3, n. 1, 2009, p. 72-84.
- LOUKIDES, M. K.. **What are ChatGPT and its friends? Opportunities, costs, and risks for large language models**. Massachusetts: O'Reilly Media, 2023.
- LUGER, G. F. **Inteligência Artificial**. São Paulo: Pearson Education do Brasil, 2013.
- LUGER, George F. *et al.* **Artificial Intelligence: Structures and Strategies for Complex Problem Solving**, 5/e. India: Pearson Education India, 1998.
- LUGER, George F. **Knowing our world**. New York: Springer International Publishing AG, 2021.

LYU, Chenyang; XU, Jitao; WANG, Longyue. New trends in machine translation using large language models: Case examples with chatgpt. [S.l.] **arXiv preprint arXiv:2305.01181**, 2023.

MAHR, Sabine. The impact of ISO/TC 37 standards on technical communication. **Revista Digital Translation**, Amsterdam/Philadelphia, v. 10, n. 2, p. 180-199, 2023.

MAIENBORN, Claudia; HEUSINGER, Klaus; PORTNER, Paul (Ed.). **Semantics-Theories**. Walter de Gruyter GmbH & Co KG, 2019.

MALUCELLI, Andreia; BAX, Marcello. Preface. Recife. **Proceedings** [...]. Recife, CEUR-WS.org, 2012. Disponível em: <https://researchr.org/publication/ontobras-2012>

MARITACA AI. MariTalk. [S.l.]: Maritaca AI, 2023. Disponível em: <https://github.com/maritaca-ai/maritalk-api>. Acesso em: out. de 2024.

MARTIN, Sean; SZEKELY, Ben; ALLEMANG, Dean. **The Rise of the Knowledge Graph**. Massachusetts: O'Reilly Media, Incorporated, 2021.

MARTINS, Ana Maria; CARRILHO, Ernestina (Ed.). **Manual de linguística portuguesa**. Alemanha: Walter de Gruyter GmbH & Co KG, 2016.

MARZINOTTO, Gabriel. FrameNet annotations alignment using attention-based machine translation. Marseille. **Proceedings** [...]. Marseille, European Language Resources Association, 2020. p. 41-47.

MATENTZOGLU, Nicolas *et al.* MapperGPT: Large Language Models for Linking and Mapping Entities. [S.l.] **arXiv preprint arXiv:2310.03666**, 2023.

MCENERY, T.; HARDIE, A. **Corpus Linguistics**. Cambridge: Cambridge University Press, 2012.

MCSHANE, Marjorie; NIRENBURG, Sergei. **Linguistics for the Age of AI**. Massachusetts: MIT Press, 2021.

MENDES, A. **Linguística de Corpus e outros usos do corpus em Linguística**. In: MARTINS, A.M.; CARRILHO, E.. **Manual de Linguística Portuguesa**. Berlin/Boston: Walter de Gruyter, 2016, 224-251.

MEYER, Lars-Peter *et al.* Llm-assisted knowledge graph engineering: Experiments with chatgpt. Wiesbaden. **Proceedings** [...]. Wiesbaden, Springer Fachmedien Wiesbaden, 2023. p. 103-115.

MIKHAILOV, Mikhail; COOPER, Robert. **Corpus linguistics for translation and contrastive studies: A guide for research**. United Kingdom: Routledge, 2016.

MINNEMA, Gosse *et al.* Frame semantics for social NLP in Italian: Analyzing responsibility framing in femicide news reports. Kyiv. **Proceedings** [...]. Kyiv, CEUR-WS, 2021. p. 1-8.

MINSKY, M. **A framework for representing knowledge**. In.: WRIGHT, F. H. G.; GORIN, R. E. *Artificial Intelligence Memo*, n. 306. Cambridge, MA: Massachusetts Institute of Technology, 1974.

MITKOV, R. **The Oxford Handbook of Computational Linguistics**. United Kingdom: Oxford University Press, 2022.

MOREIRA, Alexandra; SALOMÃO, Maria Margarida Martins. Análise ontológica aplicada ao desenvolvimento de frames. **Revista Alfa: Revista de Linguística**, São José do Rio Preto, v. 56, p. 491-521, 2012.

MOU, Xinyi *et al.* A two stage adaptation framework for frame detection via prompt learning. Republic of Korea. **Proceedings** [...]. Republic of Korea, International Committee on Computational Linguistics, 2022, p. 2968-2978.

MOUSSELY-SERGIEH, Hatem; GUREVYCH, Iryna. Enriching wikidata with frame semantics. San Diego. **Proceedings** [...]. San Diego, Association for Computational Linguistics, 2016. p. 29-34.

MU, Yongyu *et al.* Augmenting large language model translators via translation memories. [S.l.] **arXiv preprint arXiv:2305.17367**, 2023.

MÜLLER, C. **Princípios metodológicos para a construção de uma ontologia baseada na Semântica de Frames**. 2015. Tese (Doutorado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos, São Leopoldo, 2015.

MÜLLER, Carolina; CHISHMAN, Rove. Frames: uma forma de representação de conceitos em ontologias. **Revista Gatilho**, Juiz de Fora, v. 16, 2013.

MUNN, Katherine; SMITH, Barry (Ed.). **Applied ontology: An introduction**. Alemanha: Walter de Gruyter, 2013.

MUSEN, M.A. **The Protégé project: A look back and a look forward**. *AI Matters*. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.

NARAYANAN, Sridhar. Bridging text and knowledge with frames. Baltimore. **Proceedings** [...]. Baltimore, Association for Computational Linguistics, 2014.

NELSON, M. **Building a written corpus: What are the basics?** In: O'KEEFFEE, A.; MCCARTHY, M. *The Routledge Handbook of Corpus Linguistics*. London and New York: Routledge, 2010.

NIRENBURG, Sergei; RASKIN, Victor. **Ontological semantics**. Massachusetts: MIT Press, 2004.

Norma ISO 1087. **Terminology work and terminology science — Vocabulary**. Genebra. ISO, 2019.

Norma ISO 1149. **Layout of multilingual classified vocabularies**. Genebra. ISSO, 1969.

Norma ISO 704. **Terminology work – Principles and methods**. Genebra. ISO, 2022.

Norma ISO 860. **Terminology work - Harmonization of concepts and terms**. Genebra. ISO, 2007.

Norma ISO 919. **Guide for the preparation of classified vocabularies**. Genebra. ISO, 1969.

O'KEEFFEE, A.; MCCARTHY, M. **The Routledge Handbook of Corpus Linguistics**. London and New York: Routledge, 2010.

OLIVEIRA, S. **O atleta com deficiência no contexto paraolímpico: Uma análise dos frames que entram no jogo**. 2019. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos, São Leopoldo, 2019.

ONODY, Roberto N. **Teste de Turing e Inteligência Artificial**. In.: <https://www2.ifsc.usp.br/portal-ifsc/teste-de-turing-e-inteligencia-artificial/> 2021

OPENAI. **ChatGPT** (versão GPT-4 32k). 2024. Disponível em: <https://www.openai.com/>. Acesso em: 2 jul. 2024.

OPENAI. **ChatGPT** (versão GPT-4). 2024. Disponível em: <https://www.openai.com/>. Acesso em: 2 jul. 2024.

OWEN, Tony. Artificial Intelligence. **Revista Robotica**, Massachusetts, v. 6, n. 2, p. 165-165, 1988.

OZDEMIR, Sinan. **Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs**. Boston: Addison-Wesley Professional, 2023.

PAES, A; VIANNA, D.; RODRIGUES, J. **Modelos de Linguagem**. In.: CASELI, H.M.; NUNES, M.G.V. **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português – 2ª. Edição**. São Carlos: BPLN, 2024. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-modelos/cap-modelos-linguagem/cap-modelos-linguagem.html>

PAES, A.; FREITAS, C. **ChatGPT, MariTalk e outros agentes de conversação**. In.: CASELI, H.M.; NUNES, M.G.V. **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português – 2ª. Edição**. São Carlos: BPLN, 2024. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-interacao/cap-agentes-conversacionais/cap-agentes-conversacionais.html>

PAN, Jeff Z. *et al.* **Exploiting linked data and knowledge graphs in large organisations**. Heidelberg: Springer, 2017.



PAN, Jeff Z. *et al.* Large language models and knowledge graphs: Opportunities and challenges. [S.l.] **arXiv preprint arXiv:2308.06374**, 2023.

PAQUOT, Magali; GRIES, Stefan Th (Ed.). **A practical handbook of corpus linguistics**. Berlin: Springer Nature, 2021.

PARDO, César *et al.* An ontology for the harmonization of multiple standards and models. **Computer Standards & Interfaces**, [S.l.], v. 34, n. 1, p. 48-59, 2012.

PAVEL, S.; NOLET, D. **Handbook of terminology**. Canadá: Translation Bureau: 2001.

PERNA, Cristina Lopes; DELGADO, Heloísa Koch; FINATTO, Maria José. **Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa**. Porto Alegre: EDIPUCRS, 2010.

PETRUCK, M. **Frame Semantics**. Berkeley: University of California, 2001.

PICHT, H. Terminologia – Uma área de conhecimento trans e interdisciplinar. A evolução desde Eugene Wüster. Tradução Leonardo Zilio. **Revista Cadernos de Tradução, Questões de Lexicografia, Terminologia e Tradução**, Porto Alegre, n20, jan/jun, 2007.

PICHT, Heribert *et al.* Terminologia: uma área de conhecimento trans e interdisciplinar: a evolução desde Eugen Wüster. **Revista Cadernos de tradução**, Porto Alegre, N. 20 (jan./jun. 2007), p. 7-33, 2007.

PICKBRENNER, M.; ZILIO, L. **O papel das linguagens especializadas desde meados do século XX**. In.: FINATTO, M.J.; ZILIO, L. Textos e Termos por Lothar Hoffmann. Um convite para o estudo das linguagens técnico-científicas. Porto Alegre: Gráfica e Editora Palloti, 2015.

PICKLER, Maria Elisa Valentim. Web Semântica: ontologias como ferramentas de representação do conhecimento. **Revista Perspectivas em Ciência da Informação**, Belo Horizonte, v. 12, p. 65-83, 2007.

PINARBASI, Fatih; TASKIRAN, M. Nurdan (Ed.). **Natural Language Processing for Global and Local Business**. Pensilvânia: IGI Global, 2020.

PITCH, H. The science of terminology: history and evolution. **Revista Terminologija**, Lithuania, 6-26, 2011.

POLI, Roberto; HEALY, Michael; KAMEAS, Achilles. **Theory and applications of ontology: Computer applications**. Netherlands: Springer Netherlands, 2010.

RAMIREZ, Mario *et al.* Transformers, Tables and Frame Semantics. Danvers. **Proceedings [...]**. Danvers, Conference Publishing Services, 2023. p. 155-160.

RAUNAK, Vikas *et al.* Do gpts produce less literal translations?. [S.l.] **arXiv preprint arXiv:2305.16806**, 2023.

REIMERINK, A.; FABER, P. **Ecolexicon**: A frame-based knowledge base for the environment. European conference of the Czech Presidency of the Council of the EU TOWARDS ENVIRONMENT Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe, edited by Hřebíček, J., Mírovský, J.H.a.P.a., Pillmann, W., Holoubek, I. & Bandholtz, T., pages 25-27. Brno: Masaryk University, 2009.

REPPEN, R. **Building a corpus**: what are the key considerations? In.: O'KEEFFEE, A.; MCCARTHY, M. The Routledge Handbook of Corpus Linguistics. London and New York: Routledge, 2010.

RIEMER, Nick. **Introducing semantics**. Cambridge: Cambridge University Press, 2010.

ROSCH, E. Cognitive representations of semantic categories. **Journal of Experimental Psychology: General**, Romania, Vol. 104, No.3, 1975, p.192-233.

ROSCH, E. Natural Categories. **Cognitive Psychology**, [S.l.], Vol.4, No.3, 1973, p.328-350.

RÖSIGER, Ina *et al.* Acquisition of semantic relations between terms: how far can we get with standard NLP tools?. Osaka. **Proceedings** [...]. Osaka, The COLING 2016 Organizing Committee, 2016, p. 41-51.

RUPPENHOFER, J.; ELLSWORTH, M.; PETRUCK, M. R. L.; JOHNSON, C. R.; SCHEFFCZYK, J. **Change of temperature**. Berkeley: International Computer Science Institute, 2010. Available at: <http://framenet.icsi.berkeley.edu/>

RUPPENHOFER, J.; ELLSWORTH, M.; PETRUCK, M. R. L.; JOHNSON, C. R.; SCHEFFCZYK, J. FrameNet II: Extended Theory and Practice. **International Computer Science Institute**, Berkeley, 2010. Disponivel em: <http://framenet.icsi.berkeley.edu/>

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial intelligence**: a modern approach. United Kingdom: Pearson, 2022.

SABOURET, Nicolas. **Understanding artificial intelligence**. London: Chapman and Hall/CRC, 2020.

SAFAVI, Tara; KOUTRA, Danai. Relational world knowledge representation in contextual language models: A review. [S.l.] **arXiv preprint arXiv:2104.05837**, 2021.

SANCHÉZ, Maribel T. The cognitive dynamics of terminological variation. **Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication**, [S.l.], v. 17, n. 2, p. 181-197, 2011.

SASANO, Ryohei. Cross-lingual Linking of Automatically Constructed Frames and FrameNet. Marseille. **Proceedings** [...]. Marseille, European Language Resources Association, 2022, p. 6620-6625.

SAUSSURE, Ferdinand. **Curso de Linguística Geral**. São Paulo: Cultrix, 2012, 28 ed..

Schalley, A. C.; Zaefferer, D. **Ontolinguistics – an outline**. In: SCHALLEY, A. C.; ZAEFFERRER, D. (eds.). *Ontolinguistics*. Berlin: De Gruyter, 2007.

SCHALLEY, Andrea C. Ontologies and ontological methods in linguistics. **Revista Language and Linguistics Compass**, [S./], v. 13, n. 11, p. e12356, 2019.

SCOPE. In: ISO/TC 37. [Genebra: International Organization for Standardization], 7 abr. 2022. Disponível em: <https://www.iso.org/committee/48104.html>

SEGUNDA, S.T. **Harmonização da terminologia nos documentos produzidos no Gabinete de Intercâmbio do Ministério da Cultura de Angola**. 2017. Dissertação (Mestrado em Terminologia e Gestão de Informação de Especialidade) – Programa de Pós-Graduação em Linguística, Universidade Nova de Lisboa, Lisboa, 2017.

SHETH, Amit; PADHEE, Swati; GYRARD, Amelie. Knowledge Graphs and Knowledge Networks: The Story in Brief. **IEEE Computer Society**, [S./], v.23, no. 4, July/August, p. 67-75, 2019.

SHUM, KaShun; DIAO, Shizhe; ZHANG, Tong. Automatic prompt augmentation and selection with chain-of-thought from labeled data. [S./] **arXiv preprint arXiv:2302.12822**, 2023.

SIN-WAI, C. **The development of translation technology**. In: SIN-WAI, C. *The Routledge Encyclopedia of Translation Technology*. New York: Routledge, 2015.

SINGH, J. **Natural Language Processing in the Real-World Text Processing, Analytics, and Classification**. New York: Taylor and Francis Group, 2023.

SIU, Sai Cheong. Chatgpt and GPT-4 for professional translators: Exploring the potential of large language models in translation. **SSRN Electronic Journal**, New York, 2023.

SKANSI, Sandro. **Introduction to Deep Learning: from logical calculus to artificial intelligence**. Berlin: Springer, 2018.

SOUZA, Diego Spader de. **Entre conceitos e conce (p) tos: uma proposta teórico-metodológica na interface entre a onomasiologia, a lexicografia e a semântica lexical cognitiva**. Tese de doutorado (Doutorado em Linguística Aplicada). Universidade do Vale do Rio dos Sinos, 2019.

SOWA, J. F. **Knowledge Representation: Logical, Philosophical, and Computational Foundations**. California: Thomsom Learning, 2000.

SRINIVASA-DESIKAN, Bhargav. **Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras.** United Kingdom: Packt Publishing Ltd, 2018.

STEFANOWITSCH, Anatol (Ed.). **Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis.** Berlin: Mouton de Gruyter, 2006.

STEFANOWITSCH, Anatol. **Cognitive linguistics meets the corpus. Expanding Cognitive Linguistic Horizons.** Amsterdam and Philadelphia: John Benjamins, 2011.

STEFANOWITSCH, Anatol. **Corpus linguistics: A guide to the methodology.** Berlin: Language Science Press, 2020.

STRAWSON, P. F. **Individuals: An Essay in Descriptive Metaphysics.** London: Routledge, 1959.

SUN, Jiashuo *et al.* Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. [S.l.] **arXiv preprint arXiv:2307.07697**, 2023.

TALMY, L. **Cognitive Semantics: An overview.** In: MAIENBORN, Claudia; HEUSINGER, Klaus; PORTNER, Paul (Ed.). *Semantics-Theories.* Berlin: Walter de Gruyter GmbH & Co KG, 2019.

TAN, He; KALIYAPERUMAL, Rajaram; BENIS, Nirupama. Building frame-based corpus on the basis of ontological domain knowledge. Portland. **Proceedings [...].** Portland, BMC Bioinformatics, 2011. p. 74-82.

TAO, Zhengwei *et al.* Eveval: A comprehensive evaluation of event semantics for large language models. [S.l.] **arXiv preprint arXiv:2305.15268**, 2023.

TAULLI, T. **Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business.** California: Apress, 2023.

TAYLOR, J.; LITTLEMORE, J. **The Bloomsbury Companion to Cognitive Linguistics.** New York: Bloomsbury Academia, 2014.

TEMMERMAN, R. **Towards new ways of Terminology Description: The sociocognitive approach.** Philadelphia: John Benjamins Publishing Company, 2000.

TEMMERMAN, Rita; VAN CAMPENHOUDT, Marc (Ed.). **Dynamics and terminology: an interdisciplinary perspective on monolingual and multilingual culture-bound communication.** Amsterdam/Philadelphia: John Benjamins Publishing Company, 2014.

TEMMERMAN, T. Teoria Sociocognitiva da terminologia. **Revista Cadernos de Tradução**, Porto Alegre, n17, out-dez, 2004, 31-50.

TRISTÃO, A.M.D.; FACHIN, G.R.B.; ALARCON, O.E.. Sistema de classificação facetada e tesouros: instrumentos para organização do conhecimento. **Ci. Inf.**, Brasília, v.33, n.2, p. 161-171, maio/ago., 2004.

TROJAHN, Cássia; QUARESMA, Paulo; VIEIRA, Renata. A framework for multilingual ontology mapping. Morocco. **Proceedings** [...]. Morocco: LREC, 2008.

TOM, Taulli. **Artificial Intelligence Basics: A Non-Technical Introduction**. Monrovia, CA, USA: Appres, 2019.

TORRENT, Tiago Timponi *et al.* Lutma: A frame-making tool for collaborative FrameNet development. [S.l.] **arXiv preprint arXiv:2205.11840**, 2022.

TURPIN, Miles *et al.* Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. **Revista Advances in Neural Information Processing Systems**, Massachusetts, v. 36, 2024.

URBANOWICZ, Ryan J. *et al.* A semi-automated term harmonization pipeline applied to pulmonary arterial hypertension clinical trials. **Revista Methods of information in medicine**, Alemanha, v. 61, n. 01/02, p. 003-010, 2022.

UREÑA GÓMEZ-MORENO, J.M.; FABER, P.; BUENDÍA CASTRO, M. Frame blending in specialized language: Harmful algal bloom. **Revista Terminology**, Amsterdam/Philadelphia, 19(2):175-201, 2013.

UREÑA, J.M.; BUENDÍA CASTRO, M. **Semantic and Conceptual Aspects of Volcano Verb Collocates within the Natural Disaster Domain: a frame-based terminology approach**. In: GRYGIEL, M. *Cognitive Approaches to Specialist Languages*. Newcastle-upon Tyne: Cambridge Scholars Publishing, 2017, pages 330-350.

UREÑA, J.M.; FABER, P.. What the analysis of extended meaning of terms can reveal about verb semantic frame structure. **Revista Australian Journal of Linguistics**, Australia, 40 (1), 1-21, 2020.

VAJJALA, S.; MAJUMDER, B.; GUPTA, A.; SURANA, H. **Practical Natural Language Processing: A comprehensive Guide to Building Real-World NLP Systems**. California: O'Reilly Books, 2020.

VIEIRA, Renata; LOPES, Lucelene. **Processamento de linguagem natural e o tratamento computacional de linguagens científicas**. In.: PERNA, C.R.; DELGADO, H.K.; FINATTO, M.J. *Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa*. Porto Alegre: EDIPUCRS, p. 183, 2010.

VIEIRA, R. *et al.* **Web semântica: ontologias, lógica de descrição e inferência**. Web e Multimídia: desafios e soluções. Porto Alegre: SBC, p. 127-167, 2005

VILAR, David *et al.* Prompting palm for translation: Assessing strategies and performance. [S.l.] **arXiv preprint arXiv:2211.09102**, 2022.

WANG, Boshi *et al.* Towards understanding chain-of-thought prompting: An empirical study of what matters. [S.l.] **arXiv preprint arXiv:2212.10001**, 2022.

WANG, Chenguang; LIU, Xiao; SONG, Dawn. Language models are open knowledge graphs. [S.l.] **arXiv preprint arXiv:2010.11967**, 2020.

WANI, M. Arif, *et al.* **Advances in deep learning**. Berlin: Springer, 2020.

WAQUIL, M. L. ; CORTINA, A. ; PARAGUASSU, L. . **Terminologia**. 1. ed. Porto Alegre: SAGAH, 2017. v. 1. 160p .

WEI, Jason *et al.* Chain-of-thought prompting elicits reasoning in large language models. **Revista Advances in neural information processing systems**, Massachusetts, v. 35, p. 24824-24837, 2022.

WEISSWEILER, Leonie *et al.* The better your syntax, the better your semantics? Probing pretrained language models for the English comparative correlative. [S.l.] **arXiv preprint arXiv:2210.13181**, 2022.

WELLER, Katrin. **Knowledge representation in the social semantic web**. Berlin: De Gruyter Saur, 2010.

WEN, Xu; TAYLOR, John R. (Ed.). **The Routledge handbook of cognitive linguistics**. London; New York, NY: Routledge, 2021.

WENDT, Igor S.; VIEIRA, Renata. Extração de Contextos Definitórios a partir de Textos em Língua Portuguesa (Extraction of Defining Contexts from Texts in Portuguese)[in Portuguese]. Cuiabá. **Proceedings [...]**. Cuiabá, STIL, 2011.

WU, Guoxiang; YUAN, Yulin. **Lexical Ontological Semantics**. London e New York: Routledge, 2019.

WU, Yangjian; HU, Gang. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. Singapore. **Proceedings [...]**. Singapore: Association for Computational Linguistics, 2023, p. 166-169.

WÜSTER, Eugen. Die allgemeine terminologielehre—ein grenzgebiet zwischen sprachwissenschaft, logik, ontologie, informatik und den sachwissenschaften. *Revista Linguistics*, 1974.

WUSTER, Eugen. **The machine tool**: an interlingual dictionary of basic concepts comprising an alphabetical dictionary and a classified vocabulary with definitions and illustrations. London: Technical Press, 1968.

YANG, Linyao *et al.* Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling. [S.l.] **arXiv preprint arXiv:2306.11489**, 2023.

YAO, Binwei *et al.* Empowering LLM-based machine translation with cultural awareness. [S.l.] **arXiv preprint arXiv:2305.14328**, 2023.

YAO, Shunyu *et al.* Tree of thoughts: Deliberate problem solving with large language models. **Revista Advances in Neural Information Processing Systems**, New York, v. 36, p. 11809-11822, 2024.

YONG, Zheng-Xin, *et al.* Frame Shift Prediction. [S.l.] **arXiv preprint arXiv:2201.01837**, 2022.

ZHANG, Wen *et al.* Interaction embeddings for prediction and explanation in knowledge graphs. New York. **Proceedings** [...]. New York: Association for Computing Machinery, 2019. p. 96-104.

ZHANG, Xiang *et al.* Don't Trust ChatGPT when Your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. [S.l.] **arXiv preprint arXiv:2305.16339**, 2023.

ZHANG, Y; TENG, Z. **Natural Language Processing: A Machine Learning Perspective**. United Kingdom: Cambridge University Press, 2021.

ZHANG, Yichi *et al.* Making large language models perform better in knowledge graph completion. [S.l.] **arXiv preprint arXiv:2310.06671**, 2023.

ZHENG, Ce *et al.* A double-graph based framework for frame semantic parsing. [S.l.] **arXiv preprint arXiv:2206.09158**, 2022.

ZHOU, Yongchao *et al.* Large language models are human-level prompt engineers. [S.l.] **arXiv preprint arXiv:2211.01910**, 2022.

ZHOU, Yulin *et al.* Revisiting Automated Prompting: Are We Actually Doing Better?. [S.l.] **arXiv preprint arXiv:2304.03609**, 2023.

ZHU, Wenhao *et al.* Extrapolating large language models to non-english by aligning languages. [S.l.] **arXiv preprint arXiv:2308.04948**, 2023.

ZHU, Yuqi *et al.* Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. **World Wide Web**, v. 27, n. 5, p. 58, 2023.

ZILIO, Leonardo. **Terminologia Textual e Linguística de Corpus**: estudo em parceria. In.: PERNA, C.R.; DELGADO, H.K.; FINATTO, M.J. Linguagens especializadas em corpora: modos de dizer e interfaces de pesquisa. Porto Alegre: EDIPUCRS, p. 128-151, 2010.