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

MIKAELA LUZIA MARTINS

THE LEXICON AS A POSSIBILITY:
the Contribution of Semantic-Terminological Information to Lexical Substitution Tasks in Natural Language Processing

São Leopoldo
2023

MIKAELA LUZIA MARTINS

THE LEXICON AS A POSSIBILITY:
the Contribution of Semantic-Terminological Information to Lexical Substitution Tasks in Natural Language Processing

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

Orientador: Prof. Dr. Sandro José Rigo

Coorientadora: Profa. Dra. Cátia de Azevedo Fronza

São Leopoldo

2023

M386l Martins, Mikaela Luzia.
The lexicon as a possibility : the contribution of semantic-terminological information to lexical substitution tasks in natural language processing / Mikaela Luzia Martins. – 2023.
163 f. : il. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Linguística Aplicada, 2023.
"Orientador: Prof. Dr. Sandro José Rigo
Coorientadora: Profa. Dra. Cátia de Azevedo Fronza"

1. Processamento de linguagem natural. 2. Semântica de frames. 3. Semântica lexical. 4. Substituição lexical. 5. Terminologia. I. Título.

CDU 81'33

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Silvana Dornelles Studzinski – CRB 10/2524)

MIKAELA LUZIA MARTINS

THE LEXICON AS A POSSIBILITY:
the Contribution of Semantic-Terminological Information to Lexical Substitution Tasks in Natural Language Processing

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

Aprovado em (dia) (mês) (ano)

BANCA EXAMINADORA

Larissa Moreira Brangel – UFRGS

Rafael Kunst – UNISINOS

Sandro José Rigo (Orientador) – UNISINOS

Cátia de Azevedo Fronza (Coorientadora) – UNISINOS

AKNOWLEDGMENTS TO CAPES

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

To Rove, whose shoulders I was lucky to stand on
for so long.

Para Mara, Cata, Pancho y Vale con cariño.

AKNOWLEDGEMENTS

I would like to begin by thanking CAPES for financing this study and conceding me a scholarship, so I was able to pursue my master's degree.

Secondly, I appreciate my family members. Adroaldo, Ana Maria, Samuel, and Cândida, thank you for your love and support. Thank you for believing that my plans for the future will work out – even when I do not believe them myself. Thank you for encouraging me to keep going even when it seems like I have reached a wall. Thank you, Violeta and Amora, for always finding creative ways to put an authentic smile on my face. As I constantly state, no other dogs know so much about Linguistics as these two do.

Next, I deeply thank Rove Chishman for her encouragement, her support, her feedback, and especially for the endless doors she has opened not only for me, but for many other students and researchers interested in Linguistics. I would not have become a Master in Linguistics without her. Back in 2017 I was welcomed into SemanTec group with warmness and encouragement. I hope I have lived up to your expectations and I hope this piece of work makes you proud. You are a huge contributor to the path I have walked so far, and I could not have asked for anyone more capable of guiding me for so long. You taught me how to be authentic, autonomous, confident in my work, and determined. Making dictionaries under your supervision made me a researcher who is detail oriented and who thinks carefully about each and every word, for words hold so much power. Thank you again not only for this, but for many other lessons and opportunities that would fill all the pages of this work.

Sandro and Cátia, thank you for accepting the challenge of orienting and supervising a study such as this one. Your contributions and your help along the way meant so much to me. I would not be able to complete this thesis without your support. I would also like to thank Caio and the professors at the Graduate Department of Applied Linguistics at UNISINOS University for their enthusiasm and fervor when teaching Linguistic oriented classes. It was an enormous pleasure to have been a student at the department and I will always look back at these years with warmth in my heart and a taste of nostalgia.

I need to mention the SemanTec members as well: Aline, Ana, Bruna, Carol, Duda, John, Sandra, and Vitória. It was a pleasure to work by your side for so many

years. I learned so much and I was constantly inspired by people who research fascinating topics and are outstanding on what they do. I always felt motivated and encouraged after out weekly meetings.

Ana, you deserve a paragraph all to yourself! Thank you for being by my side, for encouraging me to keep going when times were difficult, for being an inspiration, and for reminding me I should take time to take care for myself every once in a while. I learn so much from you and it is comforting to have someone to discuss linguistic matters constantly in the scope of our work.

The members of VLHSem Project also deserve a shoutout. Sandro, Rafael, Ana, and Eduardo, thank you for our shared work and for constantly keeping in touch with me. I also appreciate the company's team for their support and assistance throughout the development of this work.

To my friends who are always here for me, I also say thank you. Andreza Garibaldi, Bruna Rodrigues, Joana Restelli, Tales Colman, Cody Wetherelt, John McGovern, Leonardo Gil, Landon Miller, Francisco Bustamante, Catalina Reyes, Mara Gutierrez, and Valentina Schnettler; it is such a fun time being in this planet with you all!

Lastly, but not less importantly, I would like to thank Maurício for his support and encouragement during this time. Thank you for being here for me when I needed, thank you for showing up, and thank you for being good company to study, talk, and work.

The real technology - behind all our other technologies - is language. It actually creates the world our consciousness lives in.

Andrei Codrescu

What we are trying to do may be just a drop in the ocean, but the ocean would be less because of that missing drop.

Greg Mortenson

ABSTRACT

The aim of this work is to investigate the phenomenon of lexical variation in Portuguese and English in terms alignment and lexical substitution steps in Natural Language Processing (NLP) taking into account the specialized domain of retail. As a theoretical contribution, we are based on an interdisciplinary interface that considers the postulates of the areas of Computing and Linguistics. Therefore, we offer a theoretical overview of the use of semantic information in the development of NLP systems and demonstrate ways of implementing semantic information in computational lexical bases such as WordNet, FrameNet and FrameNet Brasil. With regard to Linguistics, we rely on the definitions of Murphy (2003, 2010), L'Homme (2020) and Croft & Cruse (2004) regarding the semantic relations directed to specialized terminology. We also take into account León-Araúz & Faber's (2014) classifications and inferences regarding lexical variations and translation equivalents within the scope of Terminology. Our methodology is based on the conjectures of Corpus Linguistics and relies on the use of the Sketch Engine tool to analyze the corpora in English and Portuguese that seek to represent the terminology of the domain. The pairs of terms chosen for the research exercise of the lexical substitution task are “plant” – “site” and “material” – “article”. The terminology used in the monolingual analysis stage comes from the predictions generated by three lexical substitution models: the first one takes into account the synonymy between terms, the second one considers an additional layer of information, the word embeddings, and the third one works with the aid of an additional information layer that recovers the semantic frames. The terminology used in the multilingual analysis stage comes from the corpus used and from a collection of retail terminological bases. Our monolingual analysis seeks to classify the models' predictions according to the semantic relations and results in a categorization of terms according to the definitions of terminological variation by León-Araúz & Faber (2014). The bilingual analysis, in turn, classifies the translation equivalents of the pairs of terms according to the translation problem they represent and according to the types of equivalence that were listed by León-Araúz & Faber (2014). Finally, based on analyses of a semantic-terminological nature, our results point to improvements in lexical substitution models and automatic translation models that take into account the semantic information and

the terminological classification categories in order to advance in the quality and linguistic accuracy of the results.

Keywords: Terminology; Lexical Semantics; Natural Language Processing; lexical substitution; Frame Semantics.

RESUMO

O objetivo deste trabalho é investigar o fenômeno da variação lexical em português e inglês nas etapas de alinhamento de termos e substituição lexical em Processamento de Linguagem Natural (PLN) levando em consideração o domínio especializado do varejo. Como aporte teórico, embasamo-nos em uma interface interdisciplinar que considera os postulados das áreas da Computação e da Linguística. Portanto, oferecemos um panorama teórico sobre a utilização de informação semântica no desenvolvimento de sistemas de PLN e demonstramos maneiras de implementação de informação semântica em bases lexicais computacionais como a WordNet, a FrameNet e a FrameNet Brasil. No que tange à Linguística, apoiamo-nos nas definições de Murphy (2003, 2010), L'Homme (2020) e Croft & Cruse (2004) a respeito das relações semânticas direcionadas à terminologia especializada. Também levamos em consideração as classificações e inferências de León-Araúz & Faber (2014) a respeito das variações lexicais e equivalentes de tradução no âmbito da Terminologia. Nossa metodologia apoia-se nas conjecturas da Linguística de Corpus e baseia-se na utilização da ferramenta Sketch Engine para analisar os *corpora* em inglês e português que buscam representar a terminologia do domínio. Os pares de termos escolhidos para o exercício de investigação da tarefa de substituição lexical são "*plant*" – "*site*" e "*material*" – "*article*". A terminologia utilizada na análise monolíngue provém das predições geradas por três modelos de substituição lexical: um primeiro modelo considera a sinonímia entre termos, o segundo se volta a uma camada adicional de informação, os *word embeddings,* e o terceiro modelo atua com o auxílio de uma camada de informação adicional que recupera os *frames* semânticos. A terminologia utilizada na análise multilíngue provém do corpus utilizado e de uma coleta em bases terminológicas do varejo. A análise monolíngue busca classificar as predições dos modelos de acordo com as relações semânticas e resulta em uma categorização dos termos de acordo com as definições de variação terminológica de León-Araúz & Faber (2014). A análise bilíngue, por sua vez, classifica os equivalentes de tradução dos pares de termos de acordo com o problema de tradução que representam e com os tipos de equivalência elencados por León-Araúz & Faber (2014). Por fim, a partir de análises de cunho semântico-terminológico, nossos resultados apontam para a obtenção de melhorias de modelos de substituição lexical e modelos de tradução

automática que levem em consideração a informação semântica e as categorias de classificação terminológicas com o intuito de avançar na qualidade e a precisão linguística dos resultados.

Palavras-chave: Terminologia; Semântica Lexical; Processamento de Linguagem Natural; substituição lexical; Semântica de Frames.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CLS	Computational Lexical Semantics
FE	Frame Element
LU	Lexical Unit
NLP	Natural Language Processing
WSD	Word Sense Disambiguation

TABLE OF CONTENTS

1 INTRODUCTION

The Digital Revolution began in the latter half of the 20th century and continues to the present day. Also referred to as the Third Industrial Revolution by a few authors, the Digital Revolution marks a period in which there was a turn from mechanical and electronic technology to the digital technology, with the introduction of computers, cellphones, and other devices in the lives of the global population, and especially with the rise of internet use. This period also indicates the beginning of the Information Age, since it changed and later shaped the way we store and interact with information. Hand in hand with the mentioned technological innovations, there is the development of new ways to collect and access language-related data. Not only this, but information regarding language and linguistics is now part of a large share of the global economic sector.[1]

Naturally, this moment in history opened the doors for enterprises which linked language and technology. One may think of machine translation, probably the most common example of such interface, and it is precisely due to this interest that the studies combining language and technology began to be shaped. According to Martin Kay (2004), the field of study known as Computational Linguistics may have come to life in 1949, when Warren Weaver wrote his memorandum about machine translation and its possible applications. After this, around the 60s, the term *Computational Linguistics* gained popularity until it finally became an area of study able to stand up on its own. As defined by Ruslan Mitkov (2004, p. ix), “Computational Linguistics is an interdisciplinary field concerned with the processing of language by computers”. Nowadays, there is a variety of studies one could carry out when it comes to Computational Linguistics, not only machine translation.

The intellectuals concerned with this area are called computational linguists and they have been paying special attention to the connections between language and technology. Some of them are particularly intrigued by the contributions of a linguistic theory aimed to Natural Language Processing, especially considering the exponential growth of technology, and the rise of modern devices and their demands. For instance, SemanTec (Semantics and Technology), a research group linked to the Applied Linguistics Graduate Department at UNISINOS University in Brazil, focuses

[1] Information retrieved from: https://courses.reaktor.education/en/courses/digital-revolution/the-digital-revolution/what-is-the-digital-revolution/. Access on 11/08/22.

mainly on studies which fit the connections posed by the approximations between Semantics and technology, therefore developing studies which fall in the Computational Linguistics field. The online environment has changed the way we interact with texts and the group, being aware of these differences, has been developing online dictionaries for the past eight years and has been exploring the possibilities of an unlimited space for words.

The dictionaries published by SemanTec group are structured in such a way that they reflect the real-world knowledge the users have about the subject and about words, instead of blindly following an alphabetical order. This structure resembles other online lexical ontologies which are also concerned about semantic information, although the dictionaries are centered on the sports sphere: FIELD[2] (dedicated to soccer, contemplates Portuguese, English, and Spanish), *Dicionário Olímpico*[3] (dedicated to the Olympic sports, contemplates Portuguese and English), and *Dicionário Paralímpico*[4] (dedicated to the paralympic sports, contemplates Portuguese and English). Due to the number of possibilities within Computational Linguistics, there are studies being developed in a variety of scopes when it comes to SemanTec group, from the development of an inclusive design for the dictionaries – which would make them accessible to people with disabilities – to the analysis of the visual impact that the images can have upon the user when they think of the semantic information presented by the dictionaries. Some of the researchers opted for studies regarding the translation stage of online dictionary compilation since these resources not only show variants of the lexical units in Portuguese, but also exhibit equivalents in English (and Spanish when it comes to FIELD). This study aligns with such path and focuses on lexical variation and multilinguality.

Taking this background into account, SemanTec group's translation team became part of a project called "For Fostering Text Verticalization, Term Linking and Term Harmonization with Semantic Terminological Approaches", also known as VLHSem. Said project, as the name states, focuses on semantic terminological approaches to language aiming Natural Language Processing (NLP). NLP is a field of study that belongs to Computational Linguistics, and which is multidisciplinary and multifaceted, since it benefits from studies that belong to the areas of Linguistics, Applied

[2] http://dicionariofield.com.br/langselect. Access on 11/08/22.
[3] http://www.dicionarioolimpico.com.br/. Access on 11/08/22.
[4] https://dicionarioparalimpico.com.br/. Access on 11/08/22.

Computing, Artificial Intelligence, among others. It is concerned with the theories and applications regarding language processing and machine learning. Semantic Terminological approaches, on the other hand, focus on the approximation of semantic information aimed at terminological studies for the improvement of linguistic analyses and works. This research project was inspired by the work being developed by the VLHSem group, although is not completely associated with the demands posed by VLHSem Project.

The VLHSem is an interdisciplinary group due to the fact that it is formed by both linguists and computing professionals. There is a combination of the fields of Applied Linguistics and Applied Computing in the sense that theories and methodologies from both fields are used and adapted to the demands the group works with. One example of this practice is the study that combines knowledge graphs (a concept borrowed from the Applied Computing field) and Frame Semantics (a concept borrowed from the Linguistics field, which will be explained in detail further on) in order to better represent linguistic information in NLP.

More specifically, the demands of the VLHSem Project include – but are not limited to – finding lexical substitutes for the software developed by the company sponsoring the project. The lexical substitution task is concerned with finding replacement terms that could substitute one term in one domain with a term in another domain but referring to the same concept. Thus, this task reflects specialized language used in different fields of expertise and deals with lexical variation. The domain chosen in our case is the retail domain, due to its relevance in the global market and in the company. The retail domain regards the sale of goods and services to consumers and is a worldwide practice. Therefore, our aim is to depict and analyze the phenomena of monolingual and bilingual lexical variation within a terminological and semantic approach seeking to provide linguistic assistance to the demands of NLP.

Choosing this scenario as an inspiration for this research was a decision which also took into account the nature of Applied Linguistics as a discipline. Since this study is conducted in a Graduate Program of Applied Linguistics[5], we align with the studies conducted in the program in the sense that the interdisciplinary dimension is

[5] SemanTec Group is part of the program and integrates the scope of “Text, Lexicon, and Technology”. The program obtained a grade of 6 in the 2017-2020 quadrennium, according to the evaluation of Linguistics and Literature area from CAPES. Despite an outstanding performance, it is being discontinued by institutional decision made by the University.

necessary for the characterization of a work within the Applied Linguistics field. Thus, the combination of Linguistics and Computing Science seems to be adequate for a study in this scope, especially because the program focuses on technology and interaction. Moreover, interdisciplinary studies are important in the sense that they create new connections between fields that were close or related but not convergent in the past, having a naturally innovative character.

Considering what has been mentioned so far, this study falls into the areas of Terminology Studies, Lexical Semantics[6] and Computational Linguistics and will be dedicated to the theme of monolingual and bilingual lexical variation in NLP, considering term alignment and lexical substitution, since these are the main tasks associated with terminology management and translation in NLP. The primary objective of this master's thesis is to investigate the phenomenon of lexical variation in Portuguese and English in the term alignment and lexical substitution stages in NLP. More about term alignment, lexical substitution, and NLP will be addressed in chapter two. The secondary objectives are: i) to investigate to what extent the analysis of these elements can be helpful to the process of improving NLP models aimed at lexical substitution; and ii) to analyze what types of lexical variation occur during the lexical substitution process and find out to what extent the multilingual aspect, especially equivalence, is affected by such lexical variants.

The significance of this proposal is justified because our aim is to fill the gap in studies that investigate the interface between Computational Lexical Semantics and Terminology. This study has the potential of giving visibility to semantic-terminological studies in term alignment and lexical substitution procedures in different software, in addition to contributing to deliberations which are relevant to the tasks of term alignment and lexical substitution aimed at translation in a digital environment. Additionally, this work fills a gap regarding the theoretical background, since the results and analyzes presented here can be of future use for works inserted in this same interface. Finally, the work is justified because it seeks to present a linguistically robust perspective based on semantic relations and semantic description for the term alignment and lexical substitution steps in the software

[6] We are aware that the terms "lexical semantics" and "lexical relations" can refer to two types of relations among terms and their meanings: paradigmatic relations and syntagmatic relations. Our focus in this work are the paradigmatic relations (such as polysemy, antonymy, hyponymy, etc.) and not the lexical relations considered syntagmatic ones. Thus, when we refer to lexical semantics, semantic relations, and lexical relations throughout this work, we are referring solely to the paradigmatic relations.

development industry. Our aim is to contribute to the discussions being held in the Computational Linguistic sphere regarding NLP and terminological semantics. In summary, the significance of this study lies in the fact that limited research studies have examined the interface in which we are inserted.

In order to outline our investigation path, we organized the work as follows: chapter two addresses the field of Computational Linguistics and provides an overview of existing literature on this topic. Section 2.1 is dedicated to core notions of this work, such as the concepts of NLP, Computational Lexical Semantics, term alignment, lexical substitution, and how Computational Linguistics deals with topics such as semantic relations, semantic description, and multilinguality. Section 2.2 advances toward some of the existing computational lexicons, which are applications of the theories within the Computational Linguistics scope and how these computational lexicons address the notions presented in section 2.1.

Chapter three focuses mainly on the key linguistic aspects of this study. Thus, it is dedicated to the underlying linguistic fields supporting the study, namely Semantics and Terminology and their approaches to specialized language use, additionally to their relation to multilinguality. This chapter is divided into two distinct sections: section 3.1 addresses the semantic relations we intend to analyze. A subsection numbered 3.1.1 takes these relations into consideration under the lenses of Terminology specifically. Section 3.2, on the other hand, focuses on the field of Terminology, especially the history and approaches of specialized language use. In order to cover the whole scope of our study, this section is divided into subsection 3.2.1, focused on lexical variation in specialized language, and subsection 3.2.2, which concentrates on the phenomenon of lexical equivalence regarding multilinguality in specialized language use.

Regarding the methodology used to accomplish our objectives, chapter four aims to outline how the analyses of lexical variation and lexical equivalence will take place and what the course of action of said analyses will be. This chapter is concerned with the field of *Corpus* Linguistics and the tasks related to it – e.g., *corpus* compilation, *corpus* annotation – since this is the chosen approach to analyze the data. The *corpora* selection, compilation, and structure are described and explained in detail. We also describe the tool used to work with the corpus, the methodological steps adopted in our analyses, and the lexical substitution models we used to gather the data.

Chapter five focuses on our analysis, which has two different focuses. One of them is monolingual and is addressed in section 5.1, and the other one is bilingual and is outlined in section 5.2. We summarize what the data has shown us and exemplify what has been found. Our intention in this chapter is to demonstrate how the combination of semantic and terminological approaches to language can be used to analyze the monolingual and bilingual data provided by the lexical substitution models, in order to accomplish the objectives of our work.

Lastly, chapter six offers our final considerations contemplating the analyses made and narrates what conclusions can be made regarding the analyses performed, the theoretical background chosen, and the usability of the research. We also reflect on our objectives and how we believe they were achieved. This chapter also discusses the future possibilities of work regarding the scope of our research and the interface we are inserted. Finally, we address the contributions of our study to the fields of NLP and Linguistics.

2 SEMANTICS IN COMPUTATIONAL LINGUISTICS

As Marshall McLuhan[1] puts it: “The spoken word was the first technology by which man was able to let go of his environment in order to grasp it in a new way”. McLuhan predicted the World Wide Web nearly 30 years before it was invented and he knew very well language would have an important role to play when the technologies would eventually become part of our lives, the same way he knew the power that words have when they are used to describe our reality. In his book “The Gutenberg Galaxy”, McLuhan emphasizes how what he calls communication technology affects cognitive organization, which was one of the concerns the author had regarding the use of language in technological tools.

If McLuhan’s predictions about the rise of technology and the importance of language throughout this process seemed like an ambitious and unattainable idea a few decades ago, nowadays it is not only understandable or reasonable, but it has become a significant part of our lives. Language is crucial in every step of human development, and it would not be any different regarding technology advancement. This is one of the reasons why Computational Linguistics gained popularity as the concerns about language processing continued to grow their roots in the fields intrigued by this interface.

Computational Linguistics, as previously mentioned, was coined due to the interest in machine translation. The linguists who founded the field were particularly captivated by the syntactic structures proposed by Noam Chomsky in which grammar could be approached as a deductive system suited to computer applications, according to Kay (2003). This is probably why “syntax is by far the most mature area in natural language processing” (BATES, BOBROW & WEISCHEDEL, 2006) and why there is so much work that has been done regarding this class of language analyses. Despite not being as popular as its counterpart, semantics also contributed to the development of NLP by allowing linguists to approach word meaning in knowledge domains, which helps disambiguate word senses.

[1] Herbert Marshall McLuhan, born in 1911 and deceased in 1980, was a Canadian philosopher whose work is among the foundations of media theory. He was responsible for coining the expression "the medium is the message" and the term “global village” in his book titled Understanding Media: The Extensions of Man. Information retrieved from: https://en.wikipedia.org/wiki/Marshall_McLuhan#cite_note-13. Access on 11/08/22.

Early on, computational linguists realized that semantics would have a crucial role to play in NLP. Its relevance lies in the fact that the success of NLP systems hinges on two factors: on the one hand, sufficient language coverage is possible with relatively simple semantic models, and on the other hand, the semantics of words is constrained by the semantic relations among the chosen words and by the restrictions posed by the domains in which they are inserted. (BATES, BOBROW & WEISCHEDEL, 2006). However, much of the developments of Semantics in NLP have been made regarding reduced and narrow domains, due to the polysemic nature of general lexicon. Additionally, the semantic relations and the fuzzy nature of lexical meaning has contributed to the lack of a uniform semantic representation of language when it comes to NLP. Each working linguist must establish his/her own set of knowledge domains to address semantics in a database.

Taking into account the hurdles posed by lexical semantics to the NLP sphere, the aim of this chapter is to dive into the developments and challenges of handling semantic relations in NLP while outlining the theoretical linguistic approaches to these problems and also applications which seek to implement a semantic based approach to NLP. Thus, this chapter is divided into two main sections: one that focuses on the theory behind NLP approaches to semantic relations and another one which delineates the execution stage of semantic-based NLP. Our objective is to investigate how wide-ranging lexical databases handle the difficulties we will face in our analysis.

2.1 Natural Language Processing: Semantics

First and foremost, we must understand what NLP is and why semantic approaches are relevant to NLP development. According to Yee (2015), NLP is an area of study concerned with the handling and understanding of human languages (also known as natural languages) by computers. There are two main concerns when it comes to NLP objectives: the first is to enable the communication between humans and computers through language and the second is to develop “[…] language application systems which require considerable human language abilities and linguistic knowledge.” (YEE, 2015, p. 563). In order to achieve said objectives, this area of study focuses on tools that permit the design and application of computational systems which allow textual natural language input and output, as

stated by Yee (2015). NLP had a similar rising to the one of Computational Linguistics, given that both were created based on the demands posed by Machine Translation. However, despite being closely connected to Computational Linguistics, NLP is considered to be part of it and must not be confused with being another term used to refer to Computational Linguistics as a whole.

When it comes to the challenges and the development of NLP, still according to Yee (2015, p. 563-564),

> […] processing language by computers is much more complicated than once imagined and the knowledge required is diverse and enormous. Without satisfactorily addressing the smaller and intermediate sub-problems, it is hard to make any substantial achievement on the more sophisticated and demanding tasks like translation. […] Three related factors have played a critical role in this course of evolution in pushing language technology forward: (1) the availability of large electronic corpora, (2) the rise of statistical and machine learning approaches, and (3) the fast-growing web technology.

These three factors mentioned by Yee (2015) are still relevant to the development of NLP to this day. The importance of corpus data relies on the fact that in order to understand a language, it is necessary to have access to diverse and substantial amounts of linguistic information and diversified knowledge. *Corpora* give the linguists – and any other interested intellectuals – the means to gather together and analyze a vast volume of lexical information. This is not only useful for linguistic analysis but also

> […] has given rise to an area of research on automatic lexical acquisition, aiming to acquire a variety of lexical information including domain-specific monolingual and bilingual lexicons, significant collocation, subcategorization information, semantic similarities, selectional preferences, and others. […] In addition to knowledge acquisition, large corpora are often directly used for training statistical NLP systems, as a source from which probabilities for articular linguistic phenomena are estimated with respect to the statistical language models underlying the systems (YEE, 2015, p. 564).

As Yee (2015) draws the intersections between *corpora* use and the semantic-related issues of languages, one recognizes that the perks of *corpora* usage do not stop at extracting linguistic information or training NLP systems, on the contrary. NLP deals with an intrinsic linguistic problem, no matter the task at hand: the one of ambiguity. It can happen in the lexical level and involve part-of-speech ambiguity or sense ambiguity, or it can happen in the levels of syntactic and semantic analysis. Semantic knowledge plays a crucial role in assuring that the correct and accurate

message is portrayed in the portions of text being handled. In order to deal with this issue, NLP systems developed statistical models of language analysis and machine leaning algorithms. They became dominant in the NLP sphere due to their contribution to word sense disambiguation. Although they do not solve every problem posed by the complexities of natural languages, they facilitate the work and make the path a little less obstructed for the linguists and programmers. As Yee (2015, p. 565) puts is,

> Statistical methods have an advantage with its scalability. Its general coverage regardless of the frequency or rarity of individual linguistic phenomena overcomes the severe limitation of rule-based systems, as the efforts involved in crafting the rules often confine the resulting systems to toy systems. Statistical methods remove this hurdle, although they do not necessarily model human cognitive processes.

As Yee (2015) puts it, despite the contributions posed by statistical methods used in NLP, they do not resemble the ability which human brains have to store, access, and process multiple word senses. The uniqueness of meaning construction and meaning evocation poses a challenge which is still being addressed by many NLP studies, such as GLENSKI et al. (2021), MICHALOPOULOS et al. (2022), BAI et al. (2022), and HARVILL, GIRJU & HASEGAWA-JOHNSON (2022), to name a few.

The texts used to feed the *corpora* come from a variety of sources, one of them being the internet, which has become a rich linguistic material provider. This brings us to the third aspect mentioned above, the one of the fast-growing web technologies. The use of the web as *corpora* has considerable potential especially when it comes to the analysis of less common languages and less documented languages. Furthermore, the internet allows for an immensurable quantity of information to be stored online from an equally gigantic number of sources. Therefore, although the internet makes the compilation stage of *corpora* analysis easier, there must be a filter applied to the content that is going to be used in the *corpora*. All of the sources must be checked and approved by the compilers, taking into account the necessities of said *corpora,* otherwise the researcher may end up with useless, “polluted” material for their *corpora*.

In order to get to the word sense disambiguation stage, there are some important tasks that the NLP systems need to perform. We would like to highlight three of them. The first one is tokenization, and it involves locating and setting the word boundaries in the text. Next, there is the part-of-speech tagging of said tokens, in which labels

are assigned to each word, classifying them according to their lexical category. Up next, there is the parsing stage in which the system executes an analysis of each sentence and their meanings.

Precisely at this stage, the problem of ambiguity rises once again to haunt the working linguist. Word sense disambiguation is the process of identifying the possible senses which a polysemic word has and choosing the correct sense for the context in which the words happen. When it comes to machine translation, word sense disambiguation is vital to the correct translation of a source text, for example. Satisfactory word sense disambiguation relies on an exhaustive understanding of the source text, and it includes knowledge of the semantic relations among the words in the source text. According to Yee (2015), current practices treat word sense disambiguation as a classification task. “Systems thus attempt to assign the most appropriate sense among those given in a particular sense inventory, typically some dictionary or lexical resource, to individual words in a text.” (YEE, 2015, p. 567). This is the case of the models developed by us, which will be addressed and described in chapter four.

It is known that most NLP systems cannot afford word sense disambiguation mistakes for these inaccuracies affect the entire application in a negative, unfavorable way. Therefore, one must consider the intrinsic linguistic features of words, paying special attention to semantics. This brings us to the topic of our next subsection, the field of computational lexical semantics. But before we dive into the realm of lexical semantics aimed to NLP, there are two last NLP tasks we must address in order to contemplate the full work range of this study: sentence and term alignment tasks and the lexical substitution task.

Alignment[2], as defined by Ahrenberg (2015, p. 395), corresponds to the “process of relating part of a source text to parts of a target text”. Still according to the author, this notion is related to the use of parallel *corpora*, and its purpose is to obtain relations of equivalence and/or correspondence in a translation works. The alignment task, when approached from a computational perspective, involves statistical models aimed at the modeling of alignment characteristics. Often, more than one model is

[2] We are aware of the terminological differences between “alignment” and “correspondence” in the literature. For the purposes of this work, we adopt the definition presented by Ahrenberg (2015) and use “alignment” to refer to the totality of piece of text relations.

used in this stage, and there is the need of a model combination. (AHRENBERG, 2015).

When we look at the alignment task through the lenses of Linguistics, we realize that “alignment is a prerequisite for many tasks relating to translation technologies, including statistical machine translation, terminology extraction, population of bilingual lexicons and search in translation corpora.” (AHRENBERG, 2015, p. 397). In this sense, alignment can be of many types, including text alignment, sentence alignment, or word alignment. Since our analyses focuses mainly on the latter two types of alignment, we will not turn our attention to text alignment in this chapter.

Sentence alignment, as the name leads us to believe, refers to the alignment of texts in the level of the sentence. The quality of the sentence alignment is dependent on properties of the *corpora* used. Additionally, under ideal circumstances, there must not be unclear sentence boundaries, extra sentences, untranslated sentences, reordered sentences, sentence split, or sentence aggregation. All these factors contribute (or not) to high sentence alignment accuracy (AHRENBERG, 2015).

According to Ahrenberg (2015), sentence alignment algorithms can be of four basic kinds, and they all explore statistical tendencies when the texts being aligned are compatible with what would be the ideal *corpora*. These are the four algorithms mentioned by Ahrenberg (2015):

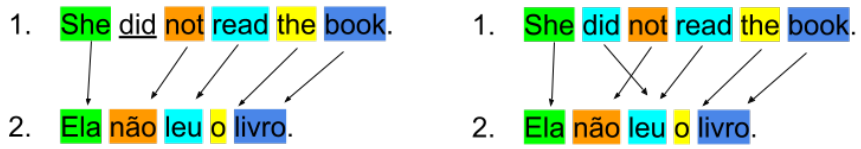
- Distribution of matches on types: the majority of the alignment is constituted of 1-1 sentence matches.
- Monotonicity: when the texts being aligned are represented as matrixes and the matches tend to occur near the diagonal of these matrixes, since the rows or columns can be used to represent the tokens or characters.
- Length: alignments can be measured by the number of characters or by the number of words, since short sentences tend to have shorter translations and long sentences tend to have longer translations.
- Token associations: alignments can be obtained by any type of association measurement, such as matching the strings.

Word alignment, on the other hand, is usually performed on sentence-aligned bitexts (AHRENBERG, 2015) and will be referred to from now on as term alignment. As stated by Ahrenberg (2015, p. 400),

> Word alignment as a computational problem is harder than sentence alignment. Many-to-many matches are more abundant, and matches may involve sets of words that are not adjacent. Moreover, null matches are generally more frequent, as is reordering. […] Word alignments are also harder to establish for humans than sentence alignments. One reason is that structure and meaning differ between languages. One language may employ prepositions to express what another uses case-endings for […].

In cases where the structure of words changes from language to language, there have been suggestions that one links the additional term to the word that has the same morphological function, whereas in other cases, the suggestions are to not align the extra terms and end up with a “null” match. Figure 1 exemplifies these morpho-syntactic differences between English and Portuguese.

Figure 1 – Possible term alignments between English and Portuguese



Source: made by the author (2023).

English traditionally uses the auxiliar verb “did” to express the simple past tense while Portuguese changes the verb form from “ler” → “leu” to express a concluded action that happened in the past. Therefore, no additional lexical units are required to express the past tense in Portuguese, which leaves the auxiliar “did” without a proper match in alignment tasks. This type of problem can be frequent and recurrent depending on the language pairs and what their morphological and syntactical characteristics look like.

As if these issues were not challenging enough, semantics also contributes to the complexity of term alignment. Translation has diverged from traditional expectations of literal translation and has been more concerned with message accuracy rather than word by word equivalence in more recent approaches. This impacts the term alignment task for semantic differences are frequently a matter of degree – and therefore a subjective decision –, instead of a precise attribute. Consequently, it becomes a hurdle to be dealt with by the alignment algorithm. When it comes to statistical tendencies for term alignment, Ahrenberg (2015) has suggested five types:

- Distribution of matches on types: at least 50% of the alignment is constituted of 1-1 sentence matches. More null matches will be included in term alignment than sentence alignment.
- Monotonicity: when languages are related the word order tends to be the same for the original text and the translation. This is true for the majority of languages, them being related or not.
- Token associations: if the terms frequently co-occur in the same order in both texts being aligned, they are likely to be a pair.
- String similarity: this type of alignment takes into account the similarities between strings, and it is more effective when the languages being paired share the same alphabet. In cases where the alphabets are different, the IPA (International Phonetic Alphabet) can be used.
- Class-based associations: when two sentences have been paired, comparisons can be made considering syntactic similarities.

The second task we must consider is the lexical substitution one. It was defined by Arefyev et al. (2020, p. 1242) as “[...] the task of generating words that can replace a given word in a given textual context.” It is a useful task for many NLP applications such as question answering, summarization, paraphrase acquisition, text simplification, and/or lexical acquisition. The elaboration of said task depends on the goal at hand and this task “[…] involves a lexical sample of nouns, verbs, adjectives, and adverbs. Both annotators and systems select one or more substitutes for the target word in the context.” (MCCARTHY & NAVIGLI, 2007, p. 48)

Thus, the lexical substitution task resembles the term alignment one due to the fact that it is also heavily affected by sense ambiguity. The task involves finding terms that are appropriate in the given context and are related to the target word in some sense. In order to achieve this,

> […] unsupervised substitution models heavily rely on distributional similarity models of words (DSMs) and language models (LMs). Probably, the most commonly used DSM is *word2vec* model (Mikolov et al., 2013). It learns word embeddings and context embeddings to be similar when they tend to occur together, resulting in similar embeddings for distributionally similar words. Contexts are either nearby words or syntactically related words (Levy and Goldberg, 2014) (AREFYEV et al. 2020, p. 1243).

This type of model is going to be used to generate the monolingual terminology we intend to analyze in our work. Thus, once the alignment task is performed and we have a list of terms to work with, the lexical substitution models work with these target terms and we provide an analysis of them to observe if the models are successfully handling sense ambiguity and other semantic related difficulties. More about it will be described in chapter four, which is dedicated to the methodology. The bilingual analysis does not depend on the models of lexical substitution because the model used by us does not provide translation substitutes, only other related words in English. More about these details will be discussed in both chapters four and five.

In our analysis further on, we intend to consider some of the factors mentioned here and lay out their impact on the final result of our alignment attempts. We intend to consider especially the lexical substitution factors that contribute to the task. Having outlined the main NLP characteristics and discussed term and sentence alignment strategies, we now turn our attention to the semantic issues related to NLP.

2.1.1 Computational Lexical Semantics

Computational Lexical Semantics is a field of study which belongs to the NLP area of Computational Linguistics and one of its main concerns is how to automate the process of representing lexical meaning of natural language expressions in computational systems and applications. According to Saint-Dizier & Viegas (1995), Computational Lexical Semantics (CLS) draws from psycholinguistics, knowledge representation, and computer algorithms and architecture. Lexical Semantics, in general, focuses on the word senses taking into consideration the nature, characteristics and relationships between these meanings. CLS describes the word senses individually, as well as their multiple relationships, however its focus is the usefulness of these descriptions for computer systems dedicated to automatically process natural language and the main concern here is not necessarily linguistic analysis or language teaching – which is one of the main preoccupations of Lexical Semantics.

Semantic information is probably the most complex and the least explored of all linguistic features of language, as we have stated before. Here lies the relevance of our study. This work is inserted in this area of computational linguistic study and

analysis, with a semantic-oriented approach. As mentioned before, due to the relevance of this interface, there are other related works which attempt to accomplish similar things, such as GLENSKI *et al.* (2021), MICHALOPOULOS *et al.* (2022), BAI *et al.* (2022), and HARVILL, GIRJU & HASEGAWA-JOHNSON (2022).

Before we approach CLS concerns in this section, however, it is worth mentioning that our aim here is to dissect semantic knowledge and semantic relations aimed at computational developments and applications. More about semantics, semantic relations, lexical semantics, lexical variation, and lexical equivalence will be presented in chapter three of this work, for chapter three focuses deeply on all the linguistic features of our analysis. Therefore, our focus here is to discuss what aspects of lexical semantics are currently being developed in Computational Linguistics, especially in NLP.

Since we are dealing directly with term alignment and lexical substitution, the most common semantic relation we will be dealing with is polysemy. In general terms, polysemy occurs when a word has more than one meaning/sense related to it (MURPHY, 2010). The opposite of polysemy is known as monosemy, which occurs when a word has a single meaning/sense related to it. In the literature review, we see that polysemy is also distinct from homonymy, which seems like an identical linguistic phenomenon but is slightly different (MURPHY, 2010). Polysemy happens when the different meanings a word has are somehow related to one another, whereas homonymy regards a coincidental similarity between two or more meanings of a word. When deciding between a polysemic or a homonymic feature of word meanings, it is common practice to look at the word etymology to check if the two (or more) word meanings are historically related or not. Lexicographers sometimes list polysemes in the same dictionary entry and list homonyms as separate headwords (MURPHY, 2010).

As stated by Alves (2009, p. 5, our translation[3]),

> The understanding of polysemy and the proposition of a computationally tractable representation for this phenomenon occupy a prominent position in

[3] Originally: “A compreensão da polissemia e a proposição de uma representação computacionalmente tratável para esse fenômeno ocupam posição de destaque na Semântica Lexical Computacional. Ninguém discute a importância de uma base de dados lexicais representar as diferentes nuances de sentido decorrentes da polissemia, mas essa é uma das poucas unanimidades relacionadas ao tema. Quando passamos a refletir sobre a melhor estratégia para a descrição desse tipo de informação, seja no âmbito da Linguística, seja no âmbito do PLN, começamos a perceber as discordâncias entre os pesquisadores.”

> the Computational Lexical Semantics. Nobody disputes the importance of a database lexical words representing the different nuances of meaning that come from polysemy, but this is one of the few unanimities related to the topic. When we begin to reflect about the best strategy to describe this type of information, whether in the context of Linguistics, whether within the scope of NLP, we begin to perceive the disagreements between the researchers.

This happens because semantics was long ignored in Linguistic studies due to its subjective character. Intellectuals such as Chomsky did not even approach semantic studies and focused on syntactic, which has culminated in the solid and constant development of syntax analysis in Computational Linguistics. With the rise of Cognitive Linguistics in the 80s leveraged by the rupture with generative models of language studies promoted (mainly but not only) by Noam Chomsky, Semantics became once again a matter of interest for linguists – only this time, it was regarded as a different object of analysis when compared to formal semantics. There was another area of study playing a part in this renewed interest in word meaning: NLP.

Considering the purposes of this chapter and the background mentioned above, we will focus mainly on the semantic relations as defined by Apresjan (1974) for two reasons. The first reason that justifies our choice is the fact that Apresjan (1974) is a much-referenced author in NLP-oriented works. The second reason concerns the organization of our study: as mentioned before, more information about the semantic relations between words will be presented in chapter three. When it comes to the semantic phenomena to be detailed, Apresjan (1974) delineates the following types of polysemy: language and speech polysemy, metonymically and metaphorically motivated polysemy, homonymy, monosemy, regular polysemy, and irregular polysemy. Considering the purpose of this work, we will approach Aprejan (1974)'s definition of homonymy, monosemy, regular polysemy, and irregular polysemy. The other polysemy types, although suitable for other types of semantic analysis, are outside of the scope of our study.

The author describes homonymy as a coincidence of two (or more) lexical units whose meanings have nothing in common (APRESJAN, 1974). Alves (2009) proposes a formula based on Apresjan (1974)'s definition in order to identify homonymy. According to the author, the senses a^1 and a^2 associated to the lexical form A are homonyms only if between them there is NO semantic coherence factor. (ALVES, 2009). Examples of it in English would be:

i) A = bat, in which case a^1 refers to the instrument used to hit a ball in a game such as baseball, and a^2 refers to the nocturnal flying mammal that inspired Batman.
ii) A = address, in which case a^1 refers to a specific location and a^2 refers to the verb meaning “to speak to someone”.
iii) A = match, in which case a^1 refers to the stick used to make a flame and a^2 refers to the action of pairing things alike.

Monosemy, on the other hand, is identified by Apresjan (1974) as the phenomenon that occurs when there is no sense duplication, and it can be seen as opposite to homonymy. What happens in this case is there is an “[…] inclusively disjunctive organization of the semantic components” (APRESJAN, 1974, p. 14) where “A = ‘B or C’, then A = 'either B, or C, or B and C at the same time’.” (APRESJAN, 1974, p. 14). An example of monosemy in English would be the term “child,” which can be used to denotate both boys and girls.

Lexical polysemy, according to Apresjan (1974, p. 14-15),

> […] will be defined through the concept of the similarity of meanings. The meanings a_i and a_j of the word A are called similar if there exist levels of semantic analysis on which their dictionary definitions (semantic trees) or associative features have a non-trivial common part. […] The word A is called polysemantic if for any two of its meanings a_i and a_j there exist meanings a^1, a^2, …, a_k, a_i such that a_i is similar to a^1, a^1 to a^2, etc., a_k to a_i and a_i to a_j. As we see, the definition does not require that there be a common part for all the meanings of a polysemantic word; it is enough that each of the meanings be linked with at least one other meaning.

Having defined the phenomenon of lexical polysemy, Apresjan (1974) moves on to the distinctions between regular and irregular polysemy. According to the author, regular polysemy occurs when polysemy of the word A with the meanings a_i and a_j in a given language, have at least one other word B with the meanings b_i and b_j being semantically distinguished from each other in exactly the same way as a_i and a_j. However, a_i and b_i, a_j and b_j are not synonyms. Examples of this type of polysemy would be:

i) A = book, in which case a_i means the physical object book (e.g., Can you pass me the book?) and a_j refers to the content of the book (e.g., I liked that book so much!) and B = handout, in which case b_i means the physical

handout (e.g., can I give you this handout?) and b_j refers to the content of the handout (e.g., this handout is not informative enough.). In this case, a_i and b_i and a_j and b_j are not synonyms.

ii) A = girl, in which case a_i means the small female child (e.g., do you see that girl playing by the pool? That is my daughter.) and a_j refers to the female partner, girlfriend (e.g., this is Tara, she is my girl!) and B = boy, in which case b_i means the small male child (e.g., that little boy is a funny kid) and b_j refers to the male friend (e.g., I met this boy when we were 20 years old). In this case, a_i and b_i and a_j and b_j are not synonyms.

Irregular polysemy, on the other hand, is described by Apresjan (1974) as what happens to A “if the semantic distinction between a_i and a_j is not exemplified in any other word of the given language” (APRESJAN, 1974, p. 16). What this means is that a lexical unit A with the meanings a^1 and a^1 presents irregular polysemy in a certain language only if there is no other lexical unit B with meanings b^1 and b^2 that are semantically distinct from one another in the exact same way as the distinction between a^1 and a^2. Furthermore, it can be said that words which possess irregular polysemy are inclined to be related to one another via metaphorical relationships. An example of irregular polysemy happens with A = position, where a^1 refers to the physical position of an object (e.g., I think the left corner is a better position for the couch) and a^2 refers to a situation or set of circumstances (e.g., the company's financial position is grim).

Additionally to having a deeper knowledge of some of the semantic relations detailed by Apresjan (1974), we must be aware that the applications which will be illustrated in the next section aim to represent the meaning of words and their relations by formally describing and organizing language. This does not equal structuring language information identically to the way in which human brains do. In fact, this task can be more accurately compared to the compilation of dictionaries and thesaurus, ontologies, and taxonomies, for example. Yet, this type of application is structured online and according to the demands of NLP, therefore, it does not suffer from the lack of space some of the mentioned lexical works do.

Computational scientists have come up with a few strategies to efficiently deal with semantics in NLP and not compromise the whole system with inadequate word sense disambiguation. One of these strategies, that considers polysemy, is the

Sense Enumeration Model, which consists of a sense listing for each word. The aim is to enumerate the totality of senses attributed to one lexical unit. Nirenburg & Raskin (2003) criticize this strategy for quite a few reasons. According to the authors, the Sense Enumeration Model lists semantic meanings but does not establish relationships between them, it is a subjective activity because it is contingent on the linguist's opinion and knowledge, and it does not predict new senses. The fact that including semantic relations in the handling of word senses is extremely beneficial to the final results is not fresh news. In fact, the success of word sense disambiguation seems to be closely related to the use of semantic relations. On the other hand, the decision to consider semantic relations in these lexical systems is closely tied to the task that NLP has in hands. As Alves (2009, p. 16, our translation[4]) puts it,

> […] for different NLP tasks, it is necessary for the system to access the lexical meaning that may be stored in different levels of generality or granularity. From there, a question that needs to be resolved is how to extract such information from lexicons and what resources are needed to perform WSD. WSD tasks are fundamentally based on automatic and semi-automatic techniques of similarity of meanings identification. The notion of similarity is used to suggest a series of sense groupings.

Before we move on to the next section, it is important to mention that the majority of WSD tasks are inserted in a multilingual context and deal with two or more languages. Not only this, but if we pay close attention to the history of NLP, most of its original tasks were concerned with machine translation, terminology management, and WSD aimed to automatic translation. Finally, term alignment and lexical substitution, other NLP tasks we will be involved with in our analysis, entail aligning terms from one language to another (or from one domain to another) and finding synonyms for target terms. Therefore, considering the multilingual background posed by the majority of NLP and AI chores, we take a bilingual approach in our work and will further discuss the implications of this decisions on chapters three, four, and five.

The objective of this section was to approach the linguistic phenomena this work is concerned with considering the perspectives of Computational Linguistics. We

[4] Originally: "[…] para diferentes tarefas de PLN, é necessário que o sistema acesse o significado lexical que pode estar armazenado em diferentes níveis de generalidade ou granularidade. A partir daí, uma questão que precisa ser resolvida é como extrair tais informações dos léxicos e que recursos são necessários para a realização de WSD. As tarefas de WSD baseiam-se fundamentalmente em técnicas automáticas e semiautomáticas de identificação de similaridade de sentidos. A noção de similaridade é empregada para sugerir uma série de agrupamentos de sentidos."

clarified some of the main topics we will be working with, which are NLP, sentence and term alignment, lexical substitution task, word sense disambiguation, and the semantic relations we must approach, paying specific attention to polysemy. The next section intends to outline applications that take into account the issues we summarized so far.

2.2 From Theory to Applications: Computational Lexicons

This section intends to depict, lay out and analyze three computational lexicons that aim to represent linguistic knowledge using NLP models and semantic and syntactic information in order to typify language information. The computational lexicons chosen to be approached in this section are Princeton's WordNet, Berkeley's FrameNet, and Federal University of Juiz de Fora's FrameNet Brasil. Our intention is not to exhaust the usability or all of the linguistic and computational information regarding each of these lexicons. Instead, we will briefly introduce each one of them and then proceed to illustrate their approaches in regard to the representation of semantic information and semantic relations.

Our aim is to present projects which are successful in their attempts to represent semantic information and semantic relations in the computational landscape. We intend to clarify that it is possible to approach semantics in NLP despite the challenges it poses. We also believe that these examples might help in the understanding of semantic information representation. We shall begin by WordNet.

2.2.1 WordNet

WordNet[5] is a large free lexical database of English language created, structured, and hosted by Princeton University. It is available for download and, as stated on the website, it is a useful tool for computational linguistics and natural language processing. Its first version, however, was built in the 80s in order to comprehend how children learn new words and it was actually perceived as a project which had a psychological character rather than a linguistic one. Throughout the years, WordNet was used by different people with different purposes. Later, it became a project concerned with NLP and gained a linguistic orientation, shaped

[5] https://wordnet.princeton.edu/. Access on 11/08/22.

mainly by the new-found uses WordNet acquired with time. Although WordNet still is preoccupied with psychological perspectives of language use, the most recent updates made to the tool are definitely an effort to include language related information.

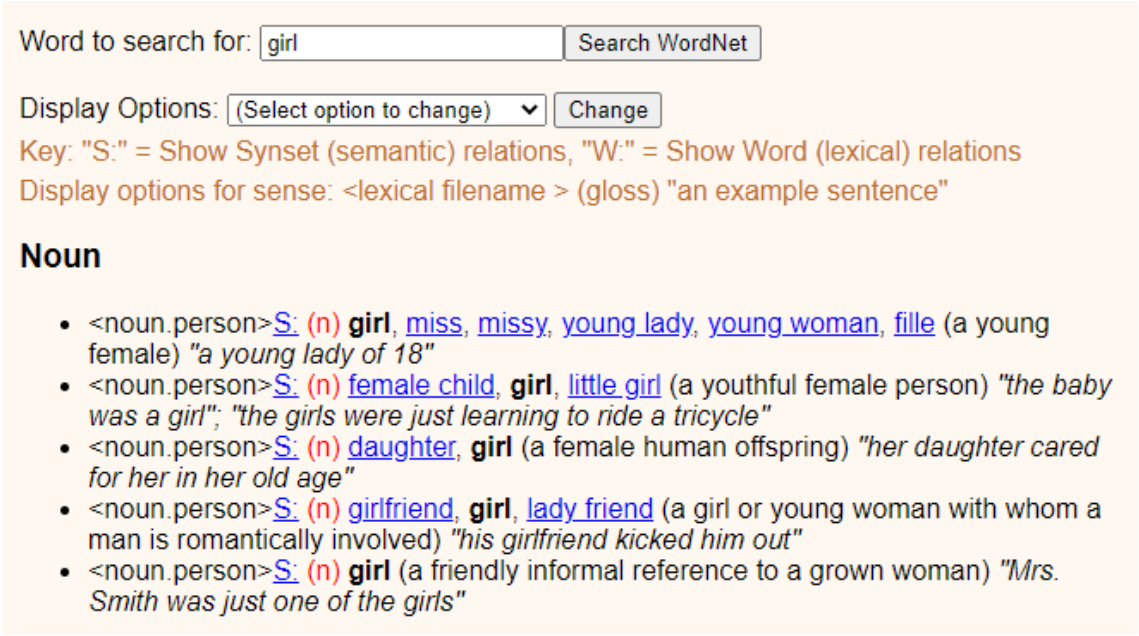
WordNet could be perceived as an online dictionary or a thesaurus, but it contains much more information other than a list of words followed by their meanings, despite the fact that it groups lexical units together based on their meanings. In WordNet nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (called synsets), each of them expressing a distinct concept. These synsets are interlinked by means of conceptual-semantic and lexical relations. The result is a network of meaningfully related words and concepts. Thus, words that are found in close proximity to one another in the network are semantically disambiguated. WordNet also labels the semantic relations among words and each of WordNet's 117.000 synsets is linked to other synsets by a small number of conceptual relations. Each synset contains a brief definition and one or more short sentences working as examples. Word forms with several distinct meanings are represented in as many distinct synsets. Thus, each form-meaning pair in WordNet is unique.[6]

When using WordNet, one types the lexical unit into the search box and defines what information the database must retrieve. Figure 2 exemplifies what results are shown by WordNet when we search for a lexical unit specifying results by selecting the "Show Lexical File Info" option. The example used here was "girl".

As we can observe from this specific search, "girl" can only be identified as a noun. Other lexical units, such as "bat", will display results classified as "nouns" and "verbs", since it can be both. WordNet uses an information label called unique beginner to introduce the lexical information. In this case, "girl" is shown under the classification <noun.person>. There are 25 unique beginners in WordNet, as Table 1 shows.

[6] Information retrieved and adapted from: https://wordnet.princeton.edu/. Access on 11/08/22.

Figure 2 – WordNet results for girl



Source: WordNet (2022).

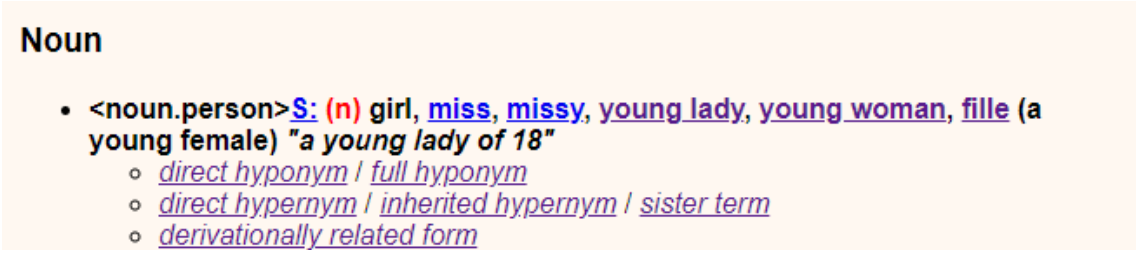
Table 1 – WordNet's unique beginners

{act, activity}	{food}	{possession}
{animal, fauna}	{group, grouping}	{process}
{artifact}	{location}	{quantity, amount}
{attribute}	{motivation, motive}	{relation}
{body}	{natural object}	{shape}
{cognition, knowledge}	{natural phenomenon}	{state}
{communication}	{person, human being}	{substance}
{event, happening}	{plant, flora}	{time}
{feeling, emotion}		

Source: Miller (1998).

Followed by the unique beginner < noun.person >, we have the option **S** which, when selected, gives us a list of semantic relations (synsets). Figure 3 exemplifies the listed synsets for "girl".

Figure 3 – WordNet synsets for girl



Source: WordNet (2022).

As we can see in Figure 3, “girl” has the following lexical relations, organized hierarchically: direct hyponym, full hyponym, direct hypernym, inherited hypernym, sister term, and derivationally related form. These are some examples of the possible relations in WordNet. There are lexical and conceptual-semantic relations which have a prestigious place in the WordNet structure and organization, as stated by Miller (1998). This is precisely what makes WordNet different from a dictionary or a thesaurus: the emphasis is in semantic content, not a list of definitions and examples. We shall briefly explain each relation in order to clarify how WordNet works, but these relations are later described in detail in chapter three. We begin by describing synonymy and antonymy in WordNet.

Synonymy is WordNet’s main relation and regards the lexical units that denote the same concept and are interchangeable in a variety of contexts. They are grouped into synsets, as already mentioned. Thus, if a word has multiple senses (displays synonymy), it will appear in multiple synsets at various locations in the taxonomy, since the nouns are organized into taxonomies where each node is a set of synonyms representing a single sense, as explained by Leacock & Chodorow (1998). The synonyms of the word “girl”, for example, are miss, missy, young lady, young woman, and fille. According to Miller (1998), WordNet’s definition of synonymy does not entail interchangeability in all contexts, but in some contexts. As outlined by Miller (1998, p. 24),

> Most synsets are accompanied by the kind of explanatory gloss that is provided in conventional dictionaries. But a synset is not equivalent to a dictionary entry. In particular, dictionary entries for polysemous words (words that can be used to express more than one meaning) have several different glosses, whereas a synset has only a single gloss. Thus, a dictionary entry can contain semantic information that, in WordNet, would be distributed over several distinct synsets, one for each meaning. It is convenient to think of a synset as representing a lexicalized concept of English. That is to say, a

> lexicalized concept is represented in WordNet by the set of synonyms that can be used (in an appropriate context) to express that concept.

Antonymy, on the other hand, refers to the definitions that could be considered opposite. According to Miller (1998, p. 39), “the strongest psycholinguistic indication that two words are antonyms is that each is given on a word association test as the most common response for each other”. Therefore, “boy” would be considered an antonym of the term “girl”.

Another relation which can be found in WordNet is hyponymy or hyperonymy. It is the most frequently encoded relation among synsets is the super-subordinate relation. It establishes relations between general items and more specific concepts. An example of this relation is the link between the terms “furniture” and “bed”. WordNet states that the category furniture includes bed, so concepts like bed (and chair, table, and desk for example) form the category furniture. All noun hierarchies ultimately go up the root node, which is one of the 25 unique beginners mentioned previously. In this case, they all go up to {entity}. Hyponymy is a transitive relation: if an armchair is a kind of chair, and if a chair is a kind of furniture, then an armchair is a kind of furniture.[7]

Meronymy is another relation worth mentioning. It is the part-whole relation between synsets in WordNet. Thus, the lexical units “fingers” and “hand” would have this type of relation, and so would “eyes” and “head”. As explained in the WordNet website, parts are inherited from their superordinates: if a chair has legs, then an armchair (which is a type of chair) has legs as well. These parts are not inherited “upward” as they may be characteristic only of specific kinds of things rather than the class as a whole. In this sense, chairs and kinds of chairs have legs, but not all kinds of furniture have legs.[8]

Table 2 shows these representations of semantic relations in WordNet and highlights some examples of each one of them. It is important to notice that these relations are among nouns, and not among other grammatical classes such as adjectives or verbs.

---

[7] Information retrieved and adapted from: https://wordnet.princeton.edu/. Access on 11/08/22.
[8] Information retrieved and adapted from: https://wordnet.princeton.edu/. Access on 11/08/22.

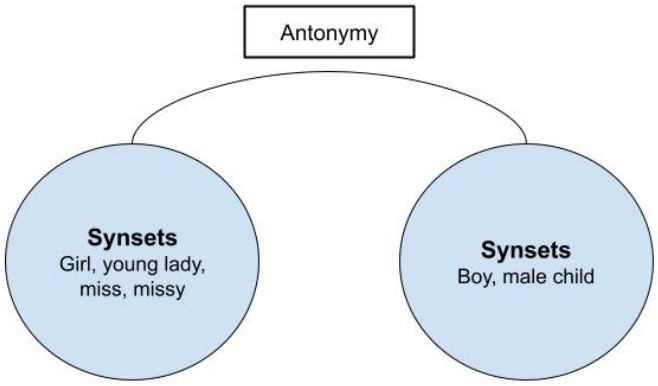
Table 2 – Semantic relations among nouns in WordNet

Semantic relation	Example
Synonymy	Girl – Miss
Antonymy	Girl – Boy
Hyponymy	Animal – Cat
Meronymy	Body – Arm

Source: made by the author (2023).

These relations are the core linguistic concepts that WordNet aims to represent. One can imagine these relations as connections that link groups of synsets in WordNet, creating webs of semantic relations and linguistic knowledge. This entire structure is an attempt to represent human knowledge about language and the concepts, referents, and experiences we conceptualize using languages as a tool. Figure 4 exemplifies this representation:

Figure 4 – WordNet's organization



Source: made by the author (2023).

This being said, WordNet is an important computational lexicon to be approached in this section because it is an application of the theories we have mentioned so far. It involves Computational Lexical Semantics, NLP, and takes advantage of studies produced in the field of Computational Linguistics. In the next section, we will analyze another computational lexicon called FrameNet.

2.2.2 FrameNet

The FrameNet[9] project is a lexical database of English language which has been in operation since 1997. It is human and machine-readable, and it is based on annotating examples of how words are used in real texts. Additionally, the data is freely available for download. It can be described as a dictionary of more than 13.000 word senses. However, just like WordNet, FrameNet also diverges from regular dictionaries in some ways. FrameNet organizes words according to their semantic information and semantic roles to help identify their meanings. As stated in the FrameNet website, the researcher in Natural Language Processing can largely benefit from the information described there, for the more than 200.000 manually annotated sentences linked to more than 1.200 semantic frames provide a unique training dataset for semantic role labeling, used in applications such as information extraction, machine translation, event recognition, sentiment analysis, among others. FrameNet-like databases have been built for other languages, including Brazilian Portuguese and we will address FrameNet Brasil once we describe the structure of FrameNet and its compromise with semantic description. There is also a new project in development that focuses on aligning the FrameNets across languages.[10] These unique features are the reason why we decided to include FrameNet in this study.

FrameNet organizes information taking into account the theory of Frame Semantics, posed by the linguist Charles Fillmore. Said theory belongs to the realm of Cognitive Linguistics, inserted in the Cognitive Semantics sphere and is based on cognitive assumptions and compromises, aiming to “provide a theoretical explanation for the relationship between language and how human beings represent situations in their minds” (L’Homme, 2020, p. 43). As stated by the linguist himself, the theory is a “research program in empirical semantics and a descriptive framework for presenting the results of such research.” (FILLMORE, 1982, p.111). In addition,

> The concept of frame does not depend on language, but as applied to language processing, the notion figures in the following way. Particular words or speech formulas, or particular grammatical choices, are associated in memory with particular frames, in such a way that exposure to a particular linguistic form in an appropriate context activates in the perceiver’s mind a particular frame – activation of the frame, by turn, enhancing access to the

[9] https://framenet.icsi.berkeley.edu/fndrupal/. Access on 11/08/22.
[10] Information retrieved and adapted from: https://framenet.icsi.berkeley.edu/fndrupal/about. Access on 11/08/22.

> other linguistic material that is associated with the frame (FILLMORE, 1976, p. 25).

Boas (2013) describes this approach as different from other theories of lexical meaning in the sense that it builds on common backgrounds of knowledge (the semantic frames) against which meanings of words are interpreted. A semantic frame, according to the theory, consists of a "cognitive structuring device, parts of which are indexed by words associated with it and used in the service of understanding." (PETRUCK, 1996, p. 2). One may also think of a frame as a system of related concepts in which in order to understand the meaning of one of the concepts, one must comprehend the whole system. The main idea of Frame Semantics was summarized as follows:

> A word's meaning can be understood only with reference to a structured background of experience, beliefs, or practices, constituting a kind of conceptual prerequisite for understanding the meaning. Speakers can be said to know the meaning of the word only by first understanding the background frames that motivate the concept that the word encodes. Whitin such an approach, words or word senses are not related to each other directly, word to word, but only by way of their links to common background frames and indications of the manner in which their meanings highlight particular elements of such frames (FILLMORE & ATKINS, 1992, p. 76-77).

This being explained, we now move on to the structure of FrameNet. As stated on the website, FrameNet shares the idea that the meanings of words can best be understood on the basis of a semantic frame, a description of a type of event, relation, or entity, and the participants in it. FrameNet gives the example of the concept of "cooking" which typically involves a person doing the cooking (Cook), the food that is being cooked (Food), something to hold the food while cooking (Container), and a source of heat for the process (Heating_instrument). In the FrameNet project, this concept is represented as a frame called Apply_heat, and the Cook, Food, Heating_instrument and Container are called frame elements (FEs). Terms that evoke this frame, such as "fry", "bake", "boil", and "broil", are called lexical units (LUs) of the Apply_heat frame.[11]

FrameNet' aim is to define the frames and to annotate sentences to show how the FEs fit syntactically around the word that evokes said semantic frame. Taking into account the Apply_heat frame, this is an example of the annotated sentence:

[11] Information retrieved and adapted from: https://framenet.icsi.berkeley.edu/fndrupal/WhatIsFrameNet. Access on 11/08/22.

... [$_{Cook}$ the boys] ... GRILL [$_{Food}$ their catches] [$_{Heating\_instrument}$ on an open fire].

In this case, the frame-evoking LU is a verb and the FEs are its syntactic dependents. In the example above, retrieved from FramNet, “boys” is the subject of the verb “grill”, “their catches” is the direct object, and “on an open fire” is a prepositional phrase modifying “grill”. The lexical entry for each LU is derived from these annotations and specifies the ways in which FEs are realized in syntactic structures headed by the word.[12]

As explained by Boas (2005), FrameNet considers an LU to be the pairing of a word with a particular sense that evokes the semantic frame. Additionally, the LU is considered the primary unit of analysis whose semantic and syntactic properties are described with respect to a semantic frame, as shown above. The example used by Boas (2005) is the term “cure” in two examples:

i) They cured the patient.

ii) They cured the pork.

As explained by the author, each sentence is linked to a different sense of cure and, thus, would evoke a different frame. The first one evokes the “cure” frame, while the second sentence evokes the “preserving” frame. What this means for FrameNet is that there are at least two distinct LUs for this verb. One of the advantages of said organization lies in the fact that it is extremely helpful for word sense disambiguation. This happens because polysemy in FrameNet is envisioned as defined by Fillmore & Atkins (1992, p. 101), when they state that “we need means of associating a word […] with particular semantic frames, and then to describe the varying ways in which the elements of the frame are given syntactic realization”. In this case, syntactic features also play an important role in word sense description.

According to Boas (2005), one of the advantages of treating polysemy by appealing to differences between semantic frames is related, as mentioned above, to WSD. For example, in cases in which an NLP application needs to determine the sense of the verb “cure”, bare semantic and syntactic information is not enough. The background knowledge provided by the semantic frames gives the application a

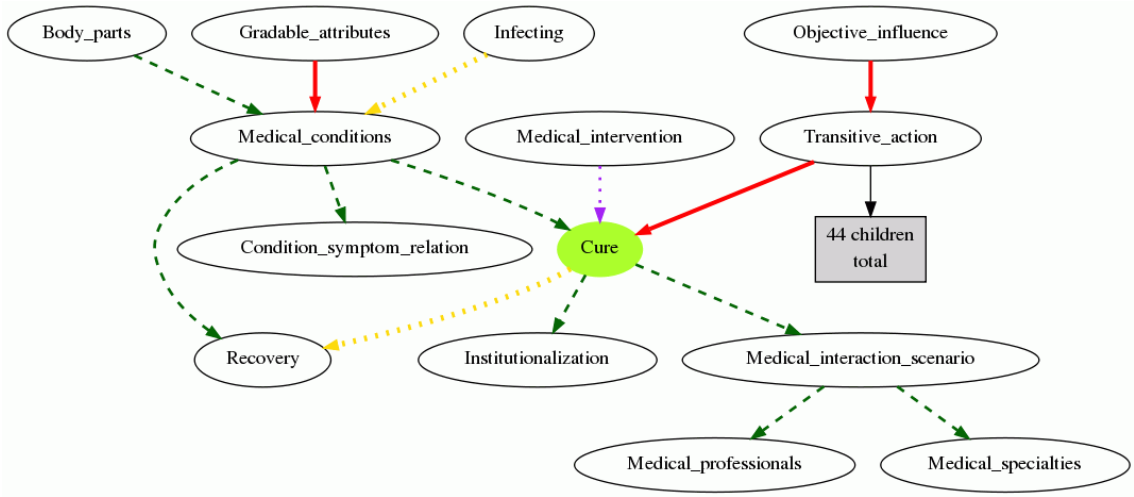
[12] Information retrieved and adapted from: https://framenet.icsi.berkeley.edu/fndrupal/WhatIsFrameNet. Access on 11/08/22.

much more reliable answer, especially useful in cases where WSD plays a crucial role in the final result, such as translation, for example.

The second advantage mentioned by Boas (2005) in this context is that it makes it possible to describe different syntactic frames occurring with the same verb as being part of the same semantic frame. This distinction appears to be determined by the fact that this sense occurs with two distinct syntactic frames, an intransitive frame (i.e., “they cured the apricots”) and a transitive one (i.e., “the apricots cured in the sun”). In FrameNet, one can observe that both the intransitive usage and the transitive one are described with respect to the same Preserving frame, thus adhering to the requirements of Fillmore & Atkins (1992, p. 101) when they say that “usage differences that need to be reported are best described, not in terms of lexical semantic differences as such, but as differences in the manner of syntactic realization of the elements of their common frame”. It is worth mentioning that this is in direct contrast to WordNet, since WordNet would provide two distinct senses for “cure”.

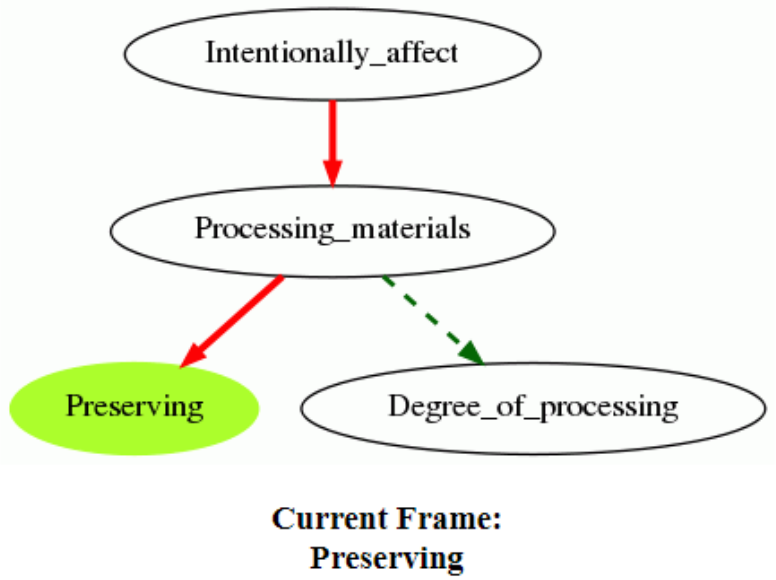
Figures 5 and 6 represent the semantic frames evoked by the verb “cure” in FrameNet and their relationships to other frames, for the purpose of a better understanding of how FrameNet organizes semantic information:

Figure 5 – Cure Frame in FrameNet.



Source: FrameNet (2022).

Figure 6 – Preserving Frame in FrameNet.



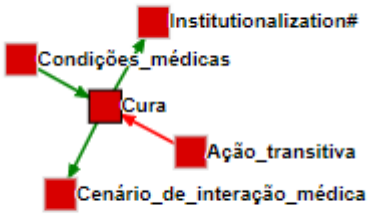
Source: FrameNet (2022).

Since our study is also based on a bilingual analysis, it is worth mentioning the FrameNet Brasil project[13], developed by researchers at the Federal University of Juiz de Fora, in Brazil. FrameNet Brasil is similar to FrameNet in the sense that the information presented is the same, and it can be viewed using similar tools and perspectives. The information is, however, in Brazilian Portuguese and the conceptualizations are organized according to the syntactic and semantic Portuguese features. FrameNet Brasil is composed of a Lexicon and a Constructicon. In similarity to FrameNet, it is also possible to visualize the relationships between the constituent elements of both through a graph. Figures 7 and 8 show the frames "*cura*" and "*preservar*" in FrameNet Brasil:

As we can observe, the relationships between semantic frames are also displayed here, as they are in FrameNet. FramNet Brasil also relies on corpus annotation to present semantic and syntactic information to the user. Additionally, FrameNet Brasil is a free of charge source of information for linguists, translators, lexicographers, among other professionals who desire to use the platform.

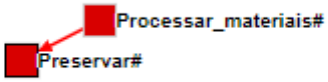
[13] https://www2.ufjf.br/framenetbr/. Access on 11/08/22.

Figure 7 – Frame "*cura*" in FrameNet Brasil.



Source: FrameNet Brasil (2022).

Figure 8 – Frame "*preservar*" in FrameNet Brasil.



Source: FrameNet Brasil (2022).

FrameNet was presented here since recent studies in linguistics (and terminology) have seen in the Frame Semantics Theory and FrameNet a "potential to characterize the specialized lexicon in a way that differs from the perspectives offered by other frameworks" (L'Homme, 2020, p. 50). It has been applied to the field of soccer (Chishman *et al*. 2014), Law (Pimentel, 2013), Computing (Ghazzawi, 2016), the environment (Faber, 2012; L'Homme, 2018), among others.

In order to close this section, we present a table designed by Boas (2005), in which the author highlights the key differences between WordNet and FrameNet. According to the linguist, because the two databases were created with distinct goals in mind, their organizational principles as well as the linguistic information presented are also different from one another.

Table 3 – Key differences between WordNet and FrameNet.

	WordNet	**FrameNet**
Theoretical background	Traditional lexical semantic relations and psycholinguistic principles	Frame Semantics
Organizational units	Words, collocations, multi-word expressions	Lexical units
Independent organizational units larger than words	n.a.	Semantic frames
Semantic relations between words	Synonymy, antonymy, polysemy, hyponymy, hypernymy, troponymy, meronymy, etc.	Polysemy, ability of a lexical unit to evoke the same semantic frame as other lexical units
Analysis of different parts of speech	In terms of different lexical hierarchies and conceptual-semantic relations	With respect to the same semantic frame
Hierarchical relations between organizational units	Multitude of different levels depending on the part of speech (e.g., troponymy, hyponymy)	Frame Inheritance, Subframe Relation, Uses Relation, 'See also' relation
Frequency information	Senses ordered by estimated frequency	n.a.
Treatment of polysemy	Influenced by syntactic properties and traditional lexicographic practice	Based on semantic frames
Syntactic information	Limited number of "sentence frames"	Exhaustive list of lexico-syntactic patterns linked to semantic information
Use of example sentences	Limited number of example sentences	Corpus example(s) for each attested lexico-syntactic pattern

Source: BOAS, 2005, p. 18-19.

The aim of this chapter was to discuss the theoretical background of Computational Semantics and the challenges and opportunities it poses for NLP. With this purpose in mind, we began by discussing Computational Lexical Semantics and other related topics such as term alignment, lexical substitution, and word sense disambiguation. These are all relevant concepts for our analysis further on. Then, we outlined two computational applications – WordNet and FrameNet – in order to exemplify how NLP projects aimed to semantic description and representation deal with lexical semantics and organize linguistic information.

The next chapter will address some of these issues taking into consideration a linguistic approach. Our intent is to depict and describe some of the semantic relations presented here under the lens of Lexical Semantics, not necessarily aimed to Computational Linguistics. Additionally, we will address some of the cross-lingual terminological and semantic implications of a bilingual analysis.

3 LEXICAL SEMANTICS AIMED AT TERMINOLOGY

It may seem odd to combine the field of Terminology, which is so well-stablished with its own methods and theorical background, with lexical semantics. However, as stated by L'Homme (2020), there has been an approximation between these fields due to the research that has been done lately – Lerat (2002a), Gaudin (2003), Faber & L'Homme (2014) to name a few – and especially due to the usability of lexical relations aimed at Terminology. L'Homme (2020) states that lexical semantics is more relevant to some of the applications of Terminology since terminological analyses need to include some degree of semantic analysis at some stage.

Lexical semantics has the potential of helping Terminology answer some questions, such as how to spot what a term is and what is not in specialized texts, which units should be included in specialized dictionaries, how to handle polysemy in specialized language, how to handle the difference between specialized and general meanings of the same term, among other issues. Thus, the combination of these fields has the potential of being fruitful and successful, especially in our case (L'HOMME, 2020).

Additionally, one could argue that semantic analysis has become a necessity in terminology studies for part of the knowledge conveyed is specialized language is reflected in the way that words behave in a text, according to Faber & L'Homme (2014). The authors also state that

> […] both general and specialized lexical items can be regarded as conceptual categories of distinct yet related meanings that exhibit typicality effects. In this regard, ontology building and conceptual modeling can benefit from the semantic analysis of linguistic concepts, based on sound theoretical principles. When terms are activated in texts, they set in motion a wide variety of underlying conceptual relations and knowledge structures. Indeed, contexts are triggering mechanisms that foreground certain relations over others (FABER & L'HOMME, 2014, p. 144).

In order to exemplify the mentioned background, the authors outline current meaning-based linguistic frameworks that can be applied to Terminology, such as Cognitive Semantics (TALMY, 2000), Frame Semantics (FILLMORE, 1977), Generative Lexicon (PUSTEJOVSKY, 1991), Lexical Grammar Model (FABER & MAIRAL, 1999), Explanatory Combinational Lexicology (MEL'ČUK et al., 1995), and so on.

Due to the mentioned factors, our work considers a semantic-terminological approach to handle the issues outlined in the previous chapter. Our aim is to semantically analyze the terminology suggested by our lexical substitution models and to terminologically analyze them as well considering term variation and lexical equivalence, since part of our analysis is bilingual. Hence, we divided this chapter in two sections which aim to discuss the relevant linguistic information necessary to our analyses. We will begin by introducing the factors related to semantics, more specifically lexical semantics, and how it connects with the terminology field. Then, we will move towards Terminology more specifically. Firstly, we will consider the monolingual approaches to Terminology, addressing lexical variation, and secondly, we move towards the bilingual one, considering lexical equivalence between languages. We believe that this division addresses the complete linguistic background which we proceed toward in our study.

3.1 Lexical Relations

This section is designed to clarify the phenomenon of lexical relations considering the linguistic perspective. In the previous chapter, we outlined Apresjan (1974)'s definitions of homonymy, monosemy, regular polysemy, and irregular polysemy due to the fact that this author is referenced in NLP-oriented works. We have also briefly described other lexical relations for the sake of explaining how WordNet and FrameNet work. However, these relations require more attention because they are central to the analysis being conducted and they are strongly tied to the problems we aim to look at. Thus, in this section we will give them more attention by describing polysemy, homonymy, hyponymy, and meronymy.

As stated by Murphy (2003, p. 3), "[…] there is no generally accepted theory of how the lexicon is internally structured and how lexical information is represented in it". Additionally, there is little if any agreement about how the conceptual information is represented or even if there is a lexical-conceptual boundary. Considering this scenario, the author identifies two types of approaches to consider when discussing semantic relations: the pragmatic one and the psycholinguistic one. In summary, Murphy (2003) does not consider semantic relations a matter of analytic or objective truth, but a matter of language users' idiosyncratic mental representations.

Murphy (2003) adopts a view of meaning which is very much aligned with the cognitivist view of meaning. In doing so, Murphy (2003) states that the mental lexicon cannot be separated from definitional and encyclopedic meaning, and that the senses of a word cannot be divided in a list like lexicographers do in dictionaries. Additionally, the author disrupts Katz and Fodor's ideas of meaning by affirming that "[…] most of our everyday content words cannot be defined by necessary and sufficient conditions" (MURPHY, 2003, p. 17). Lastly, Murphy opposes the separation of linguistic and non-linguistic information because separating these types of information would be the same as not including the representation of full senses, so conceptual meaning would still be vital in comprehending sentences. According to Murphy (2003, p. 17) "some of this lexically represented semantic information is potentially relevant to semantic relations".

We align with the author in her cognitivist position and have chosen her definitions of lexical relations due to her compromise with Cognitive Linguistics. Murphy (2003) argues that one of the problems of approaching meaning as a list of information is the fact that when lexical items map to many different concepts, polysemy arises and there is no principled limit to a word's polysemy. In order to handle this characteristic of a lexical unit and the many others that exist, Murphy (2003) makes the following assumptions about the nature of word meaning:

i) Words are polysemous: they can be associated with more than one sense.

ii) A sense is the set of conditions on a word's denotation. Connotation is a separate matter.

iii) While some semantic information may be represented in the lexicon, senses are not represented intralexically. A sense in toto is composed from whatever semantic information is specified in a lexical entry, the information (about the denotation of the word) that the word maps to in the conceptual realm, and contextual information.

In general terms, this means that Murphy (2003) rejects the monosemy solution (which we will address later), and believes

> […] senses to be dynamic and assume that the fixed mental representations of semantic information (lexical or conceptual) allow for adaptation to the requirements of a particular context. Senses that seem basic to words are usually those that require the fewest contextual cues or lexical/conceptual

> processes and/or that refer to more prototypical exemplars of the concepts involved (MURPHY, 2003, p. 20).

Croft & Cruse (2004) also understand semantic relations by using a cognitivist lens. The authors interpret polysemy in a broader way than the traditional lexicographical description which involves distinct, established senses but they include this traditional view in their interpretation as a special, prototypical case. The authors closely resemble Murphy (2003) by stating that

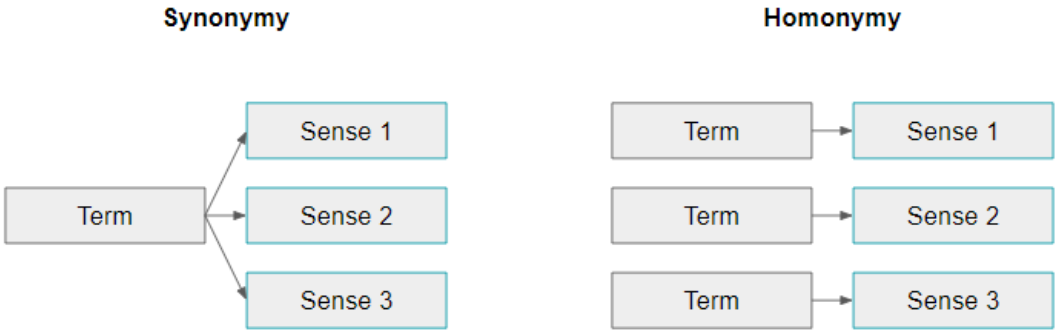
> When we retrieve a word from the mental lexicon, it does not come with a full set of ready-made sense divisions. What we get is a purport, together with a set of conventional constraints. However, in particular cases there may be powerful stable constraints favoring the construal of certain sense units. If the permanent constraints are pushing very strongly in one direction, a correspondingly strong countervailing pressure will be necessary to go against them; if the permanent constraints are weak, whether a boundary is construed or not will depend on other, mainly contextual, factors. We can portray the total meaning potential of a word as a region in conceptual space, and each individual interpretation as a point therein. Understood in this way, the meaning potential of a word is typically not a uniform continuum: the interpretations tend to cluster in groups showing different degrees of salience and cohesiveness, and between the groups there are relatively sparsely inhabited regions (CROFT & CRUSE, 2004, p. 109-110).

Having addressed the complexities of meaning, we now focus on semantic relations – which contribute to a word's meaning. Murphy (2010) identifies polysemy, homonymy, and vagueness as cases of meaning variation. As stated by the author, vagueness happens when a term has one sense that is general enough to apply to many different things. As an example, Murphy (2010) explains that "clock" means 'a device for measuring hours' but it is used to refer to a diversity of clocks, such as digital clocks and alarm clocks, for example. Thus, the term "clock" works as an example of vagueness. However, when a word represents different senses, it becomes ambiguous, as mentioned in chapter two. According to the author, we can differentiate between vagueness and ambiguity due to the fact that a vague word has an imprecise sense, while an ambiguous word has at least two senses attached to it. Murphy (2010) defines two types of lexical ambiguity: homonymy and polysemy.

Homonymy happens when the two from-meaning pairing involve different lexemes that coincidentally happen to have the same spoken/written form. Therefore, one can say that there are two lexemes that are each other's homonym. One example provided by Murphy (2010) is the term "kind" which can refer to a type of thing or an adjective that means "considerate, sweet". Polysemy, on the other hand,

is described as a single lexeme with two distinguish senses associated with it. “Book”, for example, is a term that has two meanings associated with it, one being the physical object, and the other one being the content of a book. As we can observe, these meanings are related to one another, which makes “book” different from “kind”. Figure 9 visually exemplifies the difference between homonymy and polysemy.

Figure 9 – Homonymy and polysemy



Source: made by the author (2023).

As we can observe from Figure 9, polysemy happens when one single word has multiple senses associated with it, while homonymy happens when multiple words (which share the same written and spoken form) have one single sense associated with them. Murphy (2010) also acknowledges regular polysemy, which we have briefly touched on in the previous chapter. According to the author, it happens when word senses are distinct, but which follow a general pattern or rule in the language. The author exemplifies regular polysemy by using the ‘container’ and ‘content’ senses, as identified below:

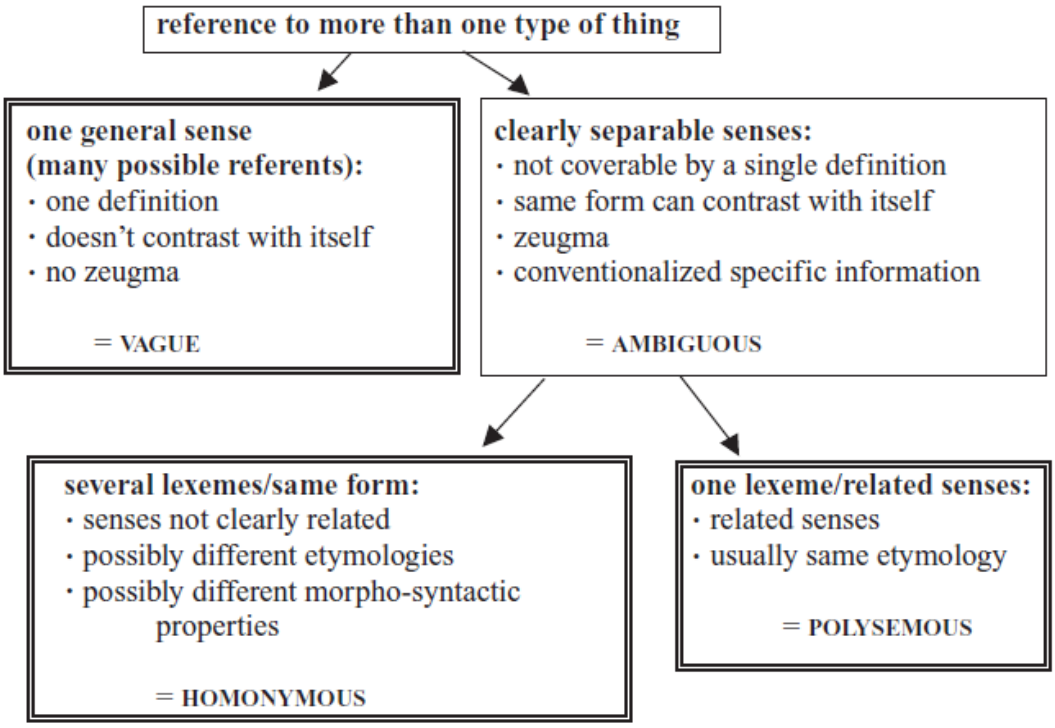
i) Container: I put some sand into a box/bottle/tin/canister.

ii) Content: I dumped the whole box/bottle/tin/canister onto the floor.

As we can observe, the relation between these senses is completely regular, therefore, predictable. As explained by Murphy (2010), if a new type of container is invented, certainly we would be able to use the name of this container to denote its contents in some situations. However, Murphy (2010) warns us that some cases of polysemy can vary in their regularity.

Lastly, Murphy (2010, p. 90) reinforces that “since vagueness, polysemy, and homonymy apply at different levels of consideration (single sense, single lexeme, different lexemes), it is sometimes the case that a single word form can illustrate all of these phenomena”. Murphy (2010) used the illustration in Figure 10 to represent each type of meaning variation.

Figure 10 – Types of meaning variation



Source: Murphy (2010, p. 91).

Thus, it is possible to clearly visualize how the author differentiates between vagueness and sense ambiguity, and between homonymy and polysemy. These are similar concepts and can be a source of confusion. Therefore, we included Murphy (2010)’s descriptions to clearly mark what these phenomena have in common and how they diverge from one another.

It is important to note that there are authors who consider meaning as a continuum, instead of a matter of three different, separate categories, and they reject the division between homonymy, polysemy, and vagueness. Tuggy (1993) is one of them. According to him, these phenomena of meaning variation should be treated as a continuum. In this sense, maximal distinctiveness would be seen as homonymy

and maximal similarity would be perceived as monosemy (characteristic of a word that has only one meaning attached to it). These ideas are also situated in the scope of Cognitive Linguistics. This approach defines the types of meaning variation outlined by Murphy (2010) by using the same types of structures and relations between them. Yet, what differentiates between them is the strength of entrenchment and semantic relatedness found in those structures.

Murphy (2010) also touches on two different approaches to polysemy, the so called monosemy position and polysemy position. In short, the monosemy position argues that there is no need to represent various senses of lexemes in the mind because each term is associated with a single, general semantic representation. This general semantic representation is, then, elaborated as a more specific interpretation according to the context in which it occurs. The polysemy position, on the other hand, holds that the different senses of a word must be separated in their representations in the mind.

The problem with the monosemy position is that despite being very explanatory in its representation of meaning, it lacks to consider the amount and range of polysemy that can be observed in natural language. It does not consider many cases of polysemy as polysemic cases because it would not be suitable for the approach. Meanwhile, the polysemy position holds that each different sense of a polyseme requires its own semantic representation and allows a deeper look at individual lexemes and their range of meanings. This monosemy solution is the one that Murphy (2003) rejects, as mentioned previously.

Because polysemy is such a complex phenomenon, there are other aspects of it to take into consideration. Words are usually considered absolute synonyms, according to Murphy (2010, p. 110), when “they are suitable in any possible context with no changes in denotation or other aspects of meaning.” In this scenario, one can affirm that there are very few cases of absolute polysemy. When it does not happen and synonym between terms is affected by the context, we can say it is a case of near-synonyms, or sense synonyms. Dialects, registers, and connotation also play a role in near-synonymy and, in this case, are variants of one another. Synonyms can be variants due to a number of non-denotational properties, including connotation, register, dialect, and affect. This definition by Murphy (2010) closely associates with the topic we handle in our analysis, thus, we decided to mention it before closing our reflections on synonymy.

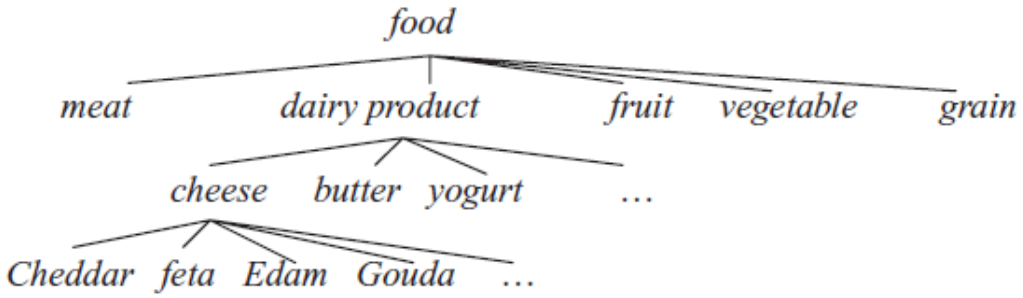
Having referred to the definitions and ways in which Murphy (2003, 2010) addresses polysemy, it is also relevant to the scope of our analysis that we address the other existing lexical relations. Due to the nature of the lexical substitution task, we will not discuss the antonymy relation (a case of contrast in which two words are opposites) since it is what we would like our models to avoid the most. Hence, we will discuss hyponymy and meronymy next.

Hyponymy refers to a type of relation that involves meanings that contain, or are contained in, other meanings. One could also argue that the extension of one word is a subset of the extension of another. Murphy (2010, p. 113) describes the relation as follows:

> […] the extension of cheddar is a subset of the extension of cheese; everything that is cheddar is also cheese, but everything that is cheese is not necessarily cheddar (since it could be gouda or mozzarella or feta instead). We could say then that cheddar is a type of cheese, and that the meaning 'cheese' is included in the meaning of cheddar.

Thus, it is fair to affirm that this relation is asymmetrical. Considering the example above, we can call "cheddar" a hyponym of "cheese" and "cheese" a hypernym of "cheddar". Figure 11 shows us examples of this relation.

Figure 11 – A partial taxonomy of food, particularly cheese



Source: Murphy (2010, p. 114).

Another characteristic of hyponymy is that it is often transitive – but not always, what makes it difficult to define. What it means is that if X is a type of Y and Y is a type of Z, then X is a type of Z. So, in this example, cheddar is a type of cheese and cheese is a type of food, hence cheddar is a type of food as well. It becomes difficult to define hyponymy because not all 'type-of' relations are transitive. For example, a specimen cup (as the one used for urine samples) is a type of cup and a

cup is a type of drinking vessel, but it is not accurate to affirm that a specimen cup is a type of drinking vessel. In order to tackle this problem, Cruse (1986) titled the proper-inclusion type of hyponymy taxonymy (since these are the relations found in classic taxonomies), while the case in which the relation is not transitive, as exemplified above, can be called functional hyponymy, since we can say 'is used as' rather than 'is a type of' in describing the relation.

Lastly, there is the issue of hyponymy being a lexical relation or not. This answer can be sought by comparing synonymy and hyponymy in terms of their behavior in language. As explained by Murphy (2010, p. 117),

> […] it is not clear that such relations are specifically lexical in nature. This is to say that the relation between the words cheese and cheddar is a direct reflection of the relation between concepts (and the objects) cheese and cheddar. The words are in a hyponym relation simply because the things that they denote are related by inclusion relations. Compare this to synonymy, for which denotative meaning is only part of the story – we noted that words are not "good" (i.e. fully substitutable) synonyms unless they match on issues like connotation, register, and dialect as well as denotative meaning. Hyponym relations are less sensitive to these non-denotational issues. It is true to say that a kitty is a type of animal, even though kitty and animal differ in register. […] For this reason, it can be said that synonymy (and, as we shall see, antonymy) is both a semantic (i.e. denotational sense) relation and a lexical (word) relation, since it involves similarity on both denotational and non-denotational levels, but hyponymy is just a semantic relation.

Having addressed the semantic relation of hyponymy, we must now look at the last relation to be discussed here, the one of meronymy. According to Murphy (2010), meronymy is a relation that considers a "part-whole" perspective. Once again we encounter an asymmetrical relation, since we can affirm that finger is a meronym of hand and hand is the holonym of finger. Another shared trait between the meronymy and hyponymy relations is the fact that meronymy also does not rely on the lexical forms of the terms because it is a reflection of the meaning of words. Murphy (2010) identifies some types of meronymy, such as:

i) **Whole > segment:** month > day
ii) **Whole > functional component:** car > engine
iii) **Collection > member:** pride > lion
iv) **Whole > substance:** pipe > copper

According to Croft & Cruse (2004, p. 159), "[…] the part-whole relation does not hold between constructed classes of elements, but between specific individuals

belonging to those classes". The authors state that there is a difference between the relation that links finger as a part of the hand, and the relation that links lake as a part of park. Because finger is always a part of a hand, Croft & Cruse (2004) call it an intrinsic construal of partness. But because when lake is a part of park it is imposed by the construal as it were from the outside, they call it extrinsic construal of partness.

Because many parts of things are optional (the tail is part of a cat, but a cat without a tail is still a cat), and because the same part-names often apply to many different wholes (e.g., "handle" is always a part-name, but is not the part of any one particular kind of thing, since doors, suitcases, and hammers all have handles), meronym relations are not as widespread or as consistent as the other relations and is generally not thought to be as central to lexical and semantic organization as the other relations. Croft & Cruse (2004) go as far as to say that there are difficulties in classifying meronymy as a lexical relation. The authors use "lid" as an example to explain that the concept of lid seems incomplete, because it has to be together with a container. However, containers do not necessarily need lids, depending on the type of container we are imagining. Thus, "lid" remains as not being as obligatory construal if partness, since not all lids are parts.

Croft & Cruse (2004, p. 162) concluded that the notion of meronymy as a lexical relation is dubious since "the implication seems to be that we cannot in general deal with the part-whole relation except at the level of the individual referent". In this same line, the authors state that

> There is, first, a very indeterminate purport, then a series of pre-meaning construals that take us nearer and nearer to the target construal and may involve a commitment to partness at some point before the final construal, but in many cases the part whole relation cannot be inferred until we reach the level of individual referents (CROFT & CRUSE, 2004, p. 162).

Taking all that was mentioned into consideration, we align with Croft & Cruse (2004) when they affirm that meronymy is a relation that applies to individual entities and is subject to variation. They also state that the recognition of this relation between construals is justified by the existence of a small number of generalizations and distinctions that only apply to this class of parts. These affirmations will be taken into consideration in our analysis further on.

Meronymy closes our thoughts and reflections on semantic relations. Since these relations can be a little confusing and they will be of great importance for our analysis further one, we included here one of Murphy (2010)'s tables, which displays a summary of the relations discussed so far (antonymy included). Figure 12 outlines the table.

Figure 12 – Properties of paradigmatic relations

	Synonym	Hyponym	Antonym	Co-hyponym	Meronym
semantic relation	similarity	inclusion	opposition	contrast	part/whole
binary	X	X	√	X	X
symmetrical	√	X	√	√	X
transitive	√	√ (taxonym)	not applicable	√	sometimes
lexical relation	√	X	often	sometimes	X

Source: Murphy (2010, p. 123).

Figure 12 summarizes the properties of each lexical relation, outlining the differences and similarities between each relation. The relations are described in terms of what they mean, if they are binary relations or not, if they are symmetrical, if they are transitive, and if they are considered a lexical relation. Thus, it provides a summary of the relations, making it easy to refer to each one of them according to the necessities of comparison.

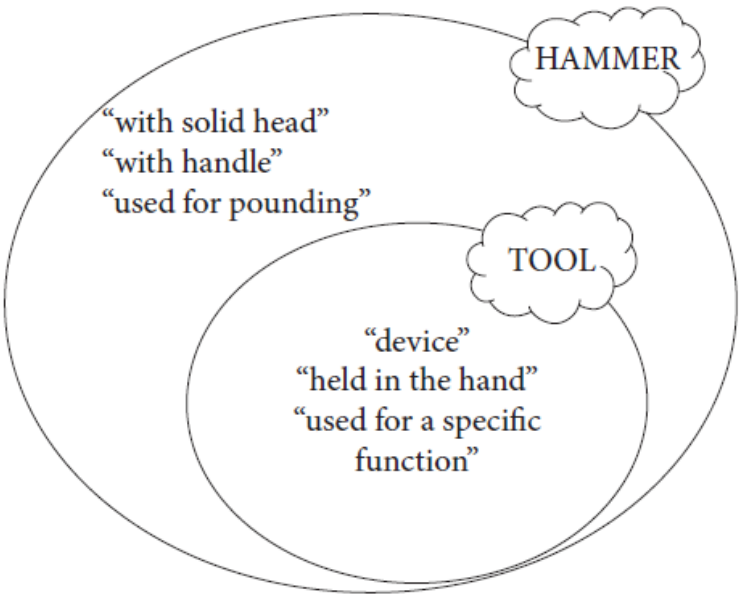
The relations described here can be accounted for Terminology. So far, we have outlined them taking a linguistic perspective into consideration but not necessarily a terminological one. Having identified each type of semantic relation that we aim to analyze, we now move on to how these relations can be used in Terminology. The next subsection addresses this issue.

3.1.1 Relations between terms in specialized language

Terminological relations, as stated by L'Homme (2020), are relations between terms and the meanings they convey. Our focus in this work are the paradigmatic relations (such as polysemy, antonymy, hyponymy, etc.) and not between other types of lexical relations such as syntagmatic ones. Thus, we will not develop any further in this second type of relation.

L'Homme (2020) identifies hypernymy and hyponymy as relations of inclusion that connect a more general term (hypernym) to a more specific one (hyponymy). Thus, her definitions closely align with the one posed by Murphy (2010). Knowledge-based approaches (which are going to be described in more detail in the next section) classify hypernyms as generic concepts and hyponyms as specific concepts. According to L'Homme (2020), these relations are fundamental to the understanding of the structure of the lexicon, since in this case they are taxonomic relations. A hypernym and a hyponym share most of their semantic content which entails that the meaning of the hypernym is included in the meaning of the hyponym. Figure 13 exemplifies these relations:

Figure 13 – Semantic components shared by hammer and tool



Source: L'Homme (2020, p. 159).

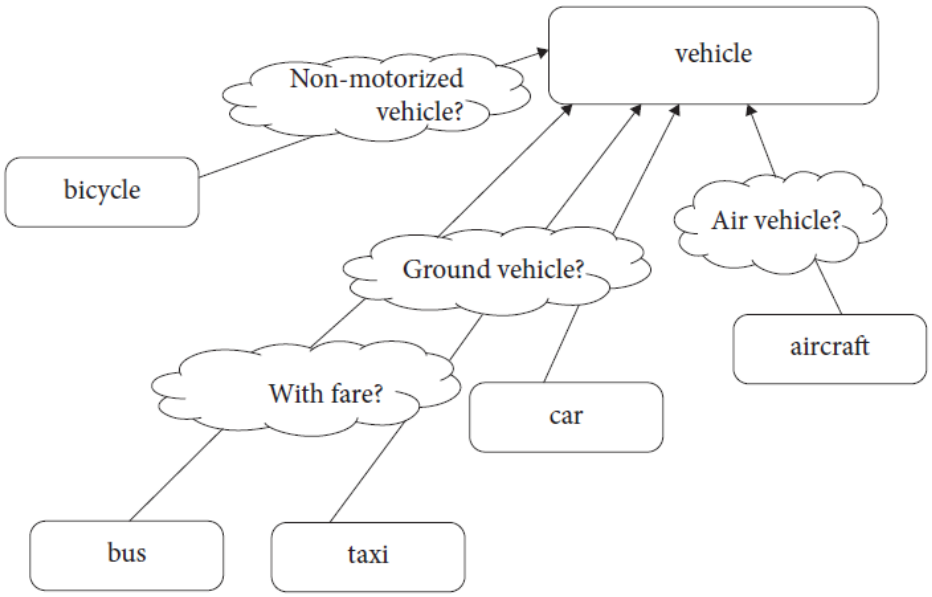
As we can observe from the example given by L'Homme (2020), the hyponym has few additional semantic components or characteristics. In this case, a tool is a device held in the hand which is used for a specific function. Hammer has the exact same information, but it is much more precise and detailed. Because we are addressing taxonomies in Terminology, the relation between both hypernyms and hyponyms is hierarchical, asymmetric, and transitive. Thus, L'Homme (2020) does not diverge from the characteristics of hyponymy and hypernymy defined by Murphy

(2010) and outlined above. But because we are dealing with natural language, there are factors that can disrupt the establishment of these relations.

> First, a lexical unit or a term might be connected to more than one hypernym depending on the semantic component that is taken into account. An obvious hypernym for the LU cat is feline. But we could also consider the cat to be a kind of pet. Of course, if we are compiling a terminological resource in any field connected directly or indirectly to biology, feline would be the adequate hypernym. However, there might be cases where pet would be a better choice, for example, if a terminological resource is concerned with different kinds of stores or the management or a city.
> Secondly, when relations are established strictly on the basis of attested LUs or terms, some gaps may appear in a hierarchy. In some cases, it might not always be possible to find names for useful sublevels (L'HOMME, 2020, p. 159).

In order to exemplify the second case, when there are gaps in the classification of items, L'Homme (2020) uses the example of types of vehicles. In this case, there is a lack of terminology in English to refer to specific types of characteristics, for example the (many) differences between ground and air means of transportation. Figure 14 outlines the gaps by using "?" after each perceived gap.

Figure 14 – Types of vehicles and lexical gaps



Source: Murphy (2020, p. 160).

As we can observe, despite being useful to Terminology, the hyponymy relation does not solve every problem or cover every issue that the specialized

lexicon may present. Unfortunately for us, L'Homme (2020) did not address meronymy in Terminology. Thus, we now move on to synonymy.

The definition of synonymy aimed at terminology studies lands in a much-referenced territory, since it is also regarded as a symmetric relation between terms that have the same meaning or very close meanings. L'Homme (2020) defines exact synonymy form a terminological perspective as what happens when two terms share all of their semantic components. Thus,

> In a pair of exact synonyms, member 1 can replace member 2 in all the sentences where member 2 appears. Conversely, member 2 should be a valid candidate to replace member 1 without affecting the meaning of sentences in which member 1 is used. Furthermore, everything that characterizes one member of the pair is also valid for the other (L'HOMME, 2020, p. 161).

Examples of exact synonyms would be the terms "hydropower" and "hydroelectricity", which can be substitutes of one another in any given context. The sentences below are examples of this type of synonymy outlined by L'Homme (2020).

Figure 15 – Examples of exact synonymy

HYDROPOWER uses a (mostly) renewable source, (water), and it long has been a favorite electricity source for many countries with large water flows.
(Bjork et al. 2011)

Waterpower, also called HYDROELECTRICITY, is a renewable form of electricity generation that harnesses the energy produced from the movement of falling or flowing water. *(Ontario Ministry of Energy and Infrastructure 2010)*
Waterpower, also called HYDROPOWER

Source: L'Homme (2020, p. 161).

However, by now we know that this is not the norm. Usually, terms share many semantic components but not all of them. Consequently, they will not be synonymous in every context. The example provided to elicit this case is the pair "habitat" and "territory". The shared semantic information:

i) they both designate specific kinds of locations.
ii) they both are used by species.
iii) they both are found in larger areas.

Despite these shared semantic traits and the possibility of substitution of one another in certain contexts, we are no longer dealing with exact synonymy. We are now looking at near-synonymy. In the examples shown by Figure 16, the replacement is not possible because “territory” is defined as a space used for a specific purpose. Although “habitat” is also a location in which species carry out different activities, the component ‘for a specific purpose” is not a core component of its meaning as it is for “territory”.

Figure 16 – Examples of “territory” and “habitat”

Broods generally remain on the nesting TERRITORY. *(Campbell 1995)*
?nesting HABITAT
Protection of breeding TERRITORIES *and nesting birds from human disturbance is also a priority. (New Mexico Department of Game and Fish 2017)*
?breeding HABITAT

Source: L’Homme (2020, p. 162).

This concludes our focus on the lexical relations applied to specialized language. As we can observe, there is not a significant difference in the definition of these core concepts. The main difference is the application, in this case since it is aimed at terminological studies and not purely semantic ones. These relations will probably appear in our data – synonymy being the main one – and we believe the theoretical background presented so far is enough for our analysis. In the next section, we intend to focus on Terminology itself and its characteristics. Additionally, we will discuss lexical variation and lexical equivalence.

3.2 Terminology: history and scope

Terminology[1] is a field of study concerned with the systematic study of the labeling and assignment of particular lexical units to one or more subjects or areas of human activity, through research and analysis of terms in context, with the aim of documenting and promoting their correct use. According to Cabré (1993, p. 37, our

[1] Considering Krieger & Finatto (2004) we adopt the term “Terminology”, written with capital ‘t’ in order to mention the field of study concerned with “terminology”, or written with the lowercase letter ‘t’, which, on the other hand, refers to the collection of words used in an area of study or discipline.

translation[2]), "[…] Terminology is the formal reflection of the conceptual organization of a specialty […]". Considering Terminology as a field concerned with conceptual precision is valuable in the sense that it ensures accuracy when it comes to lexical meaning. In areas such as Medicine and Law, for example, ambiguity in communication can be a source of problems and complications with serious consequences. Thus, using the correct terminology is crucial so conceptual exactness is achieved.

However, the path to get to this point is made of different approaches. Initially, the Terminology field considered meaning an intrinsic property of terms and did not have much importance, as explained by Faber & L'Homme (2014). Consequently, the semantics of specialized terminology was not in the spotlight and little attention was given to it. Yet, things started to change in the 90s and this change culminated in new approaches. The doors of semantic analysis were opened by descriptive terminology approaches and the rise of Corpus Linguistics. Faber & L'Homme (2014) cite some questions that these new methods of analysis raised, such as issues related to polysemy, multidimensionality, corpus pattern analysis, among others.

According to L'Homme (2020), there are different approaches to linguistic content and meaning in terminology. The first one is known as knowledge-driven, and it considers concepts as a stable structure with limits set by necessary and sufficient conditions. So, for example, every animal that fits the criteria "warm-blooded, has feathers, and lays eggs" could be considered a bird. The problem is that not all concepts can be easily delineated, and many entities usually can be classified under two or more categories. Additionally, the knowledge-driven approach considers the concepts first and the linguistic labels come second, after the concept has been identified.

Another approach described by L'Homme (2020) is the lexicon-driven approach, which is semasiological due to the fact that it takes the lexical unit as the starting point and no longer depends on the prior delimitation of concepts. One important aspect of this approach is the possibility of comparing relations between lexical units since it is relational. It can observe relations such as sameness of meaning (start - begin), opposite meaning (sustainable - unsustainable), and/or inclusion of meaning (flower - tulip). As one can imagine, this possibility opens the doors of semantic

[2] Originally: "[...] a terminologia é o reflexo formal da organização conceitual de uma especialidade [...]"

analysis when we take these lexical relations into consideration. According to L'Homme (2020), terminological analysis should incorporate aspects of both approaches in order to fully consider the linguistic content.

If one term happens to have multiple meanings and more than one is relevant in the same domain, one could use this combination of approaches to make semantic distinctions. As explained by L'Homme (2020, p. 30), "a lexicon-driven approach […] can examine the relations of each meaning conveyed by polysemous items using the meanings of other lexical units", while a knowledge-driven approach would be helpful when one needs to associate each meaning with a specific item in the knowledge structure.

Therefore, considering this information and what has been addressed in the previous section, one can argue that semantic information has become extremely important to terminology studies, even when it comes to translation. Semantics offers a rich amount of information about meaning of words that can benefit terminological choices in one or more languages because it assures conceptual precision, avoids ambiguity, and contributes to a satisfactory and successful communication.

The disambiguation of words and the selection of the correct word sense are crucial for conceptual precision, but they are not the only factors in play. Context is also essential in this regard. For instance, the word "mouse", when used in Computing Science, means the device we move around in a surface in order to move the cursor on our computer screen. It can also refer to the screen cursor itself in some cases. However, when used in the Biological and Veterinary Sciences, the same lexical unit describes a small furry animal with a long tail which belongs to the mammal family. This example may seem simplistic, but the same happens with words which are considered to be more prototypically technical, such as "ultra-high-performance concrete" or "starting hill official". Therefore, polysemous words require a clear indication of what field is being referenced, so the vocabulary can be understood precisely, and the necessary conceptual information can be activated. Context usually provides said semantic information. León-Araúz & Faber (2014, p. 2) further clarify this matter by stating that:

> We need a representation framework that allows for the inclusion of different syntactic, lexical, conceptual and semantic features, but it also needs to account for dynamism and context, which happen to influence all of these features at different levels. We understand dynamism as the changing nature of both concepts and terms due to contextual constraints, whereas

> context is defined by the different pragmatic factors that modulate such dynamism (e.g. specialized domains, cultures, communicative situations). As a consequence of their natural dynamism, concepts may be recategorized and have their relational behavior constrained, whereas terms may show several types of variants with different cognitive, semantic and usage consequences. Context is thus an important construct when describing the concepts and terms of any domain in monolingual resources […].

Context also plays an indispensable role in multilingual ventures. Undoubtedly, Terminology has always been a relevant area of study, but it has only recently gained the attention it deserves due to the fact that globalization has contributed to the rapid spread of information across not only fields of expertise, but also languages. This is the reason why a terminological approach can take into account one or more languages, thus being regarded as a monolingual, bilingual, or multilingual terminology study, depending on how many languages are involved. Therefore, Terminology can be closely linked to Translation Studies, another field of work which largely benefits from the studies and applications of Terminology. This is why we decided to divide this section in two subsections to properly address both monolingual and bilingual aspects of Terminology. Translation Studies contributes notably to the development of technologies aimed to terminological purposes since the management of lexical units is a central element of good readability and technical correctness of translated texts, for example, and it also pays close attention to contextual aspects. According to León-Araúz & Faber (2014, p.2),

> […], in multilingual resources, context also affects interlingual correspondences. When dealing with multilingual ontologies, context features must be extended to include translation relations and degrees of equivalence. As a result, a believable and useful knowledge representation needs to account for and classify context types as well as the result they may cause.

Hence, this setting makes the closeness of Translation Studies and Terminology inevitable and valuable, especially in digital ontologies. According to Krieger & Finatto (2004), the biggest motivation for this proximity between the two fields is related to the fact that "technical-scientific terms are key elements, cognitive nodules, of specialized texts." (KRIEGER & FINATTO, 2004, p. 66, our translation[3]). The authors clarify that when translators face the series of textual demands needed for a satisfying translation exercise, they understand that the correct terminology and the

[3] Originally: "termos técnico-científicos são elementos chave, nódulos cognitivos, dos textos especializados."

corresponding scientific terms consist of a unique communication form which belongs to a determined field of expertise. This is how the specialized lexicon works as a linguistic element to knowledge representation and becomes a resource to measure how successful and efficient communication in an area is.

In this line, it is extremely important that the translator responsible for the translation is acquainted with the vocabulary used – and, in order for this linguistic knowledge to be acquired, there must be conceptual knowledge as well. Therefore, the translator must know the field of expertise in question and constantly update his expertise by studying the area. This is the reason why the translator must have access to rich, complete, and well-compiled terminological repertoires in both the source and the target languages. Krieger & Finatto (2004, p. 67, our translation[4]) developed further on the importance of said knowledge:

> [...] an adequate use of terminology contributes to the achievement of semantic-conceptual precision, a condition that every specialized text translation necessarily requires. In addition to this qualification, the transposition of terms that are specific to a field from one language to another gives the translated text much of the expressive characteristics commonly used by professionals in the same field. This is also the case with the use of specialized phraseologies, which characterize typical forms of expression in professional communications. To this extent, the respect for the professional use of terms and phraseology is also a respect for style, which will favor the acceptability of the target text, regardless of the language into which it will be translated.

A good, specialized text translation must, consequently, also conform to the technical style used in writing, minimizing the characteristics of a poor translation, which would, in turn, signalize to the target text reader that this work was done by someone who is not acquainted with the material being produced in the area in at least one of the languages. In consonance to this, Krieger & Finatto (2004) propose a table with a list of technical texts characteristics and the required skills of a technical translation based on Gamero (2001). The table is organized as shown in Table 4.

---

[4] Originally: “[...] uma utilização adequada da terminologia contribui para o alcance da precisão semântico-conceitual, requisito que toda tradução de texto especializado obrigatoriamente requer. Além dessa qualificação, a transposição de uma língua para outra dos termos próprios de uma área confere ao texto traduzido grande parte da das características expressivas comumente usadas pelos profissionais do mesmo campo de atuação. Este também é o caso do uso das fraseologias especializadas, que caracterizam formas típicas de expressão das comunicações profissionais. Nessa medida, o respeito pelo uso profissional de termos e das fraseologias é também um respeito pelo estilo, o que vai favorecer a aceitabilidade do texto de chegada, independente da língua em que será traduzido.”

Table 4 - Characteristics of technical texts and skills required for technical translation

Characteristics of textual functioning → Required translator skills	
Importance of the thematic field → Knowledge of technical areas	Ability to document technical texts
Use of specific terminology → Application of appropriate technical terminology in the target language	
Presence of technical gender characteristics → Mastery of conventional features of technical genres in the target language	

Source: Krieger & Finatto (2004, p. 67, our translation).

What Table 4 displays is information regarding which skill the translator must have in order to be able to successfully handle the characteristics of technical texts. Therefore, when it comes to the importance of the thematic field, the translator must have enough knowledge of the area; when it comes to the usage of specialized terminology, one must know how to apply the terminology in the target language; and when it comes to the characteristics of this gender, there must be enough mastery of conventional features of this gender by the translator.

As one can conclude, Terminology and Translation Studies have multiple reasons to merge together and can benefit from said convergence, especially when it comes to the development of computational linguistic resources. Part of the development of said resources must account for the cognitive, contextual, and cultural aspects of language use and this is exactly where the linguists' work and knowledge can be of huge value. The attention to linguistic aspects is the core of high-quality lexical ontologies or lexical applications.

Considering what has been discussed so far, we believe that in order to successfully select the correct word when dealing with term alignment and lexical substitution in particular fields and/or choosing the correct equivalent when working with specialized language, and achieving a good and satisfying technical translation, there has to be a stage in which the linguist studies the fields and becomes aware of the conceptual knowledge and the conceptual structures that perpetuate said area.

Finally, the proximity of both fields discussed in this section of the chapter is the reason why the remaining parts of the text are divided in sections which account for lexical variation – a monolingual matter that could be related to Terminology – and lexical equivalence – a subject which belongs to the Translation Studies sphere. In the next subsections, we will consider the implications of both phenomena and how to deal with them in Computational Linguistics, especially in the NLP context.

3.2.1 Lexical Variation

In order to fulfill the objectives of this study, lexical variation is the first issue we will address, since it is approached by Terminology, and it is the phenomenon that we have been working with, considering the data made available to us during the course of the project. Lexical variation happens when there is one – or more than one – term that could be used as a substitute for a particular lexical unit. L'Homme (2020, p. 66) defines terminological variation, or term variation, as "[…] the phenomenon whereby the names of concepts change". As stated before, the field knowledge is an important requirement when it comes to a successful term selection. However, it is not the only one. We will explore the lexical variation types we may encounter in our analysis and also other issues related to defining which possibility is the best fit when it comes to term alignment and lexical substitution.

However, before we acknowledge term variants and related aspects, we must elicit the challenges that term variation poses for Terminology. According to L'Homme (2020), firstly, there is the issue of identifying which lexical units denote the same concept. Secondly, there is the issue of understanding which variant is applied with which purpose. Some of these variants, which Freixa (2006) calls denominative, are candidates for inclusion if terminological resources and databases. These variants could be confused with synonyms, but it is important to keep in mind that although denominative variants can include synonyms, they are not restricted to being synonyms. Other types of variants, called by L'Homme (2020) contextual variants, are not included in terminological resources because their purpose is to help acquire knowledge about the senses and/or the concepts involved.

Following this line of thinking and seeking to address problems in terminological variation, León-Araúz & Faber (2014) assure that the first aspect one must pay attention to is context, echoing what has already been mentioned. According to the authors, context is described

> as the parts of a written or spoken statement that precede or follow a specific word or phrase, and which can influence its meaning or effect. It is also the situation, events or information that are related to something, and which help a person to understand it. Context can have a wider or narrower scope and can include external factors (situational and cultural) as well as internal cognitive factors, all of which interact with each other. In many cases, context is the only factor that can be used for word sense disambiguation, and it also influences the choice of a word form over its variant (LEÓN-ARAÚZ & FABER, 2014, p. 5).

The authors also stated that the key to successfully select the correct term when working in the computing field is to parametrize context, in order to enable the system to be aware of situational meaning constraints. León-Araúz & Faber (2014, p. 5) also clarify that “if this is done for each language separately, cross-lingual mappings would be enhanced”. We believe that the study proposed by the authors not only takes into consideration the majority of factors related to working with term variation, but they also indicate the types of variants we could come across when working with lexical substitution in the VHLSem project. Thus, we believe it is valuable to consider their findings in the scope of this work.

León-Araúz & Faber (2014) address term dynamics by electing types of lexical variation which may affect semantics, pragmatics, and linguistic interlingual correspondences. The reason why we decided to use their definitions is the fact that these authors take into consideration the cognitive aspects of terminology and translation. Thus, their precepts will also be used further on to explain term equivalence. These variant types were defined by the authors as follows:

- Orthographic variants with no geographic origin (e.g. aesthetics, esthetics) or with geographic origin (e.g. color, colour). These variants do not affect semantics or the communicative situation, therefore, although could be found by us in our analysis, we will not further elicit them.
- Diatopic variants which include orthographic variants (e.g. groyne, groin) that do not affect semantics, dialectal variants (e.g. gasoline, petrol) which may affect semantics if culture-bound factors highlight or suppress any of the semantic features, culture-specific variants (e.g. sabkha, dry lake) which affect both semantics and the communicative situation when referring to a particular entity that, in a specific culture, adds more specific features, and calques, which may affect semantics and the communicative situation and are the result of an interlinguistic borrowing for different reasons, such as the influence of a particular language on a specialized domain. These are geographical variants which can be described as synonymous terms.
- Short form variants such as abbreviations (e.g. temp. for temperature) and acronyms (e.g. laser, Light Amplification by Stimulated Emission of Radiation). They do not affect semantics but only the communicative situation and they may be found in our data.

- Diaphasic variants which can be of three types: science-based variants, informal variants, and domain-based variants. Science-based variants, which can be scientific names (e.g. dracaena draco, drago), expert neutral variants (e.g. ocellaris clownfish, amphiprion ocellaris), jargons (e.g. in medicine, lap-appy would correspond to laparoscopic appendectomy, but no lay user would use this term), formulas (e.g. H2O, water; CaCO3, pearl), or symbols (e.g. $, dollar). These variants do not affect semantics but only the communicative situation. The scientific names refer to specialized nomenclatures and are especially useful in botany, zoology, chemistry, etc. The expert neutral variants would be the default term choice in a specialized scenario. The jargon terms are used when experts have their own informal way to refer to specialized concepts. The formulas do not affect semantics but only the communicative situation. Informal variants, on the other hand, can be lay user variants (e.g. dragon tree, drago), colloquial variants (e.g. fracking, hydraulic fracturing), or generic variants (e.g. sea, ocean; erosion, weathering). They do not necessarily affect semantics but especially the communicative situation. The lay user variants would be the default term choice in non-specialized scenarios. The generic variants are very informal variants that can activate terms pointing to different levels of conceptual granularity and thus affect semantics. Finally, there are the domain-based variants (e.g. sludge, mud) which may affect semantics and/or the communicative situation when term preferences change across specialized domains. These variants are related to what is specialized terminology and what is not. We believe that the domain-based variants are going to be the most common in our data.
- Dimensional variants (e.g. Gutenberg's discontinuity, core-mantle boundary) which are usually multi-word terms that affect semantics, since they convey different dimensions of the same concept (the person who first named it and the two parts it delimits). We do not believe this type of variant will appear in our data.
- Metonymic variants (e.g. escollera, espigón), which, as the name reveals, is based in a semantic relations. Therefore, they affect semantics because the metonymic variant designates the concept according to its parts. This type of variant will probably be very likely to happen in our analysis.

- Diachronic variants which only reflect old uses of terms. This type of variant seems unlikely to appear in our data.
- Non-recommended variants (e.g. in medicine, mental retardation has now negative connotations and has been substituted by intellectual disability) that affect connotation. This type of variant, just like the diachronic, seems unlikely to appear in our data.
- Morpho-syntactic variants (e.g. the action of the waves, wave action), which do not affect semantics but depend on collocates, term selection preferences, and the communicative situation. We believe this might appear in our analysis.

Considering the variant types described and their different consequences for communication and meaning, we believe that if a term can have numerous variants which are going to be used according to a different framework, than a work focused on term alignment must account for these linguistic features. Some of the dynamics explained above may not appear in our data. However, we can predict that a few, such as short form variants, domain-based variants and metonymic variants, for example, will be found due to the nature of the term alignment work: if we are going from one field of expertise to another, then we must expect to find conceptual, contextual, and semantic differences between terms.

The various conceptual contexts that can be found in texts involve the conceptual relations activated by the words in a sentence and their relation to the other words around them. In Terminology, this can pose a problem because it affects opaque noun compounds, and, in turn, make them more difficult to process. For instance, León-Araúz & Faber (2014) explain that when “sediment” is the head word, in noun+noun compounds the slot activated is usually (but not always) “location” (e.g. intertidal zone sediment, streambed sediment, aquifer sediment), whereas in adjective+noun compounds the “material” slot is triggered (e.g. lithogenous sediment, biogenous sediment, hydrogenous sediment, cosmogenous sediment). Therefore, the authors conclude that the analysis of heads and slots can contribute to the extraction of hyponyms, which are semantic relations between words that we expect to analyze in our data. The authors stated that

> […] it can also be useful in the study of dimensional variants that show the dynamics of concepts, where synonyms designate the same concept but add or suppress semantic slots (e.g. Gutenbergs discontinuity, core-mantle boundary). Thus, multi-word terms, whose formation is dynamic by nature, show that concepts may be classified according to multidimensional facets (location, material, etc.) and can be a rich source for semantic features modelling. However, semantic features should not always be stable in representations (LEÓN-ARAÚZ & FABER, 2014, p. 7).

As pointed out by León-Araúz et al. (2013), not all dimensions are always part of a unique conceptualization for their activation is context-dependent. There is also the factor of multidimensionality in Terminology, defined by Rogers (2004), which can cause a given concept to have two hypernyms in the same domain – due to multiple inheritance. Yet, multidimensionality is not always responsible for hypernymy relations, it can also leave the door open for non-monotonic inheritance in some cases. Much of these semantic relations happen because, according to Cimiano *et al*., (2010) changes in conceptualization and in the terminology are not independent from each other. This is the reason why domains and culture-bound characteristics of terms should be addressed when working with NLP.

Here is where the semantic relations and sense of words come into play. As stated in chapter two and in first section of this chapter, the semantic relations are an important part of our terminological analysis. Taking into consideration what León-Araúz et al. (2013) stated so far, we believe that we will be able to classify each suggested term according to one type of lexical relation – providing a semantic analysis – and according to the variants defined by the authors – providing a terminological analysis. The combination of both would result in a semantic-terminological analysis because we intend to draw connections, similarities and/or differences between one another.

With the listed situation in mind, we expect to address in our analysis stage not only the lexical variants identified in the data, but also which semantic relations appear in the lexical substitution model suggestions. The next subsection will address another terminological aspect of this study, however, with focus on multilinguality and cross-linguistic conceptualizations.

3.2.2 Lexical Equivalence

Lexical equivalence is a linguistic phenomenon associated with the search for a sameness of meaning for words, either in two or more languages. This topic has

been a source of controversy and, ironically enough, many authors have tried to come up with a satisfying definition for the term “equivalence” but there seems to be a difficulty finding common ground. In Translation Studies, a definition of equivalence is tied up with the theoretical approach chosen by each author and the purpose of the translation work. The intend of this section is not to define equivalence[5], but it is to address the types of equivalence one can come across when dealing with multilinguality in ontologies and term alignment and lexical substitution. The matter of equivalence has not been highly discussed in the field of Terminology studies as well.

Since this work focuses on term alignment and lexical substitution tasks, equivalence will be found in the lexical level. Therefore, we will not account for other types of equivalence, such as sentence equivalence or communicative equivalence, for example. We will, however, consider conceptual equivalence and term equivalence due to the nature of this work. Moreover, both are related to our understanding of what can (and should) be considered equivalence, since we closely associate with cognitivist views of meaning.

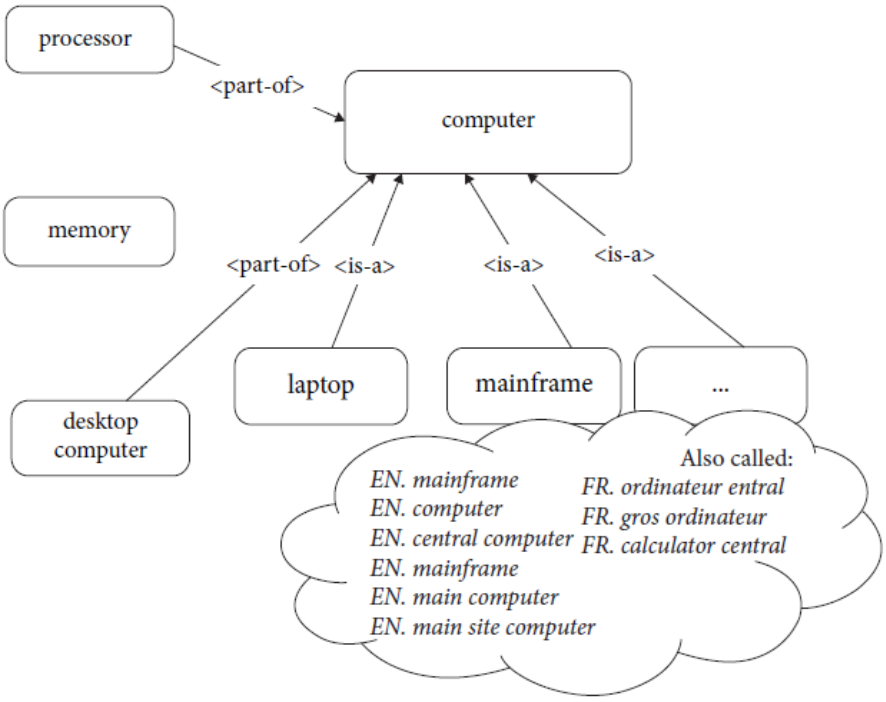
According to L’Homme (2020), conceptual equivalence is associated with knowledge-driven approaches to terminology. According to this approach, terms are considered equivalents if they belong to different languages and denote the same concept within the same domain. Thus, one can affirm that conceptual equivalence is concerned with exact equivalence, or sameness of meaning. In this case, an ontology, for example, should not suffer many alterations if we decide to add equivalents to the structure, since the concepts remain the same. Identical designations in one language are called synonyms, but identical designations in different languages are called equivalents. The example in Figure 17 shows the labels in English and French that L’Homme (2020) used to exemplify conceptual equivalence.

As we can observe in Figure 17, the semantic relations are present in this type of ontology, providing associations for equivalents and terms in the original language. Furthermore, not much importance is given to the structure of terms. Thus, the equivalents can be of many types, varying from a single term lexical unit to

[5] A panoramic view of the equivalence definitions throughout the years can be found at Martins (2019). Equivalence has been analyzed by the author regarding two fields of language work: Translation Studies and Bilingual Lexicography.

collocations and multi-word terms. However, terminological equivalence is established differently since it considers the basis of meaning of the terms.

Figure 17 – Labels in English and French in a conceptual structure



Source: L'Homme (2020, p. 231).

Lexicon-driven approaches define terminological equivalence as the relation between terms that belong to different languages and convey the same meaning in the same domain, as explained by L'Homme (2020). One example of lexical equivalence is the terms "ecosystem" in English and "*écosystème*" in French, since they have the same definition. In this case, however, polysemy heavily affects equivalence. When this is the case, there are different situations that can occur considering lexical units in different languages, according to L'Homme (2020):

i) A lexical item carries (at least) two different meanings: the first one is associated with general language and the second one is associated with a specialized domain. Example: the noun "key" is used in general language where it designates 'a small object used to open doors or boxes'. In computing, it designates 'a part of a keyboard pressed by a user to insert a character or send a command' (among others). When it

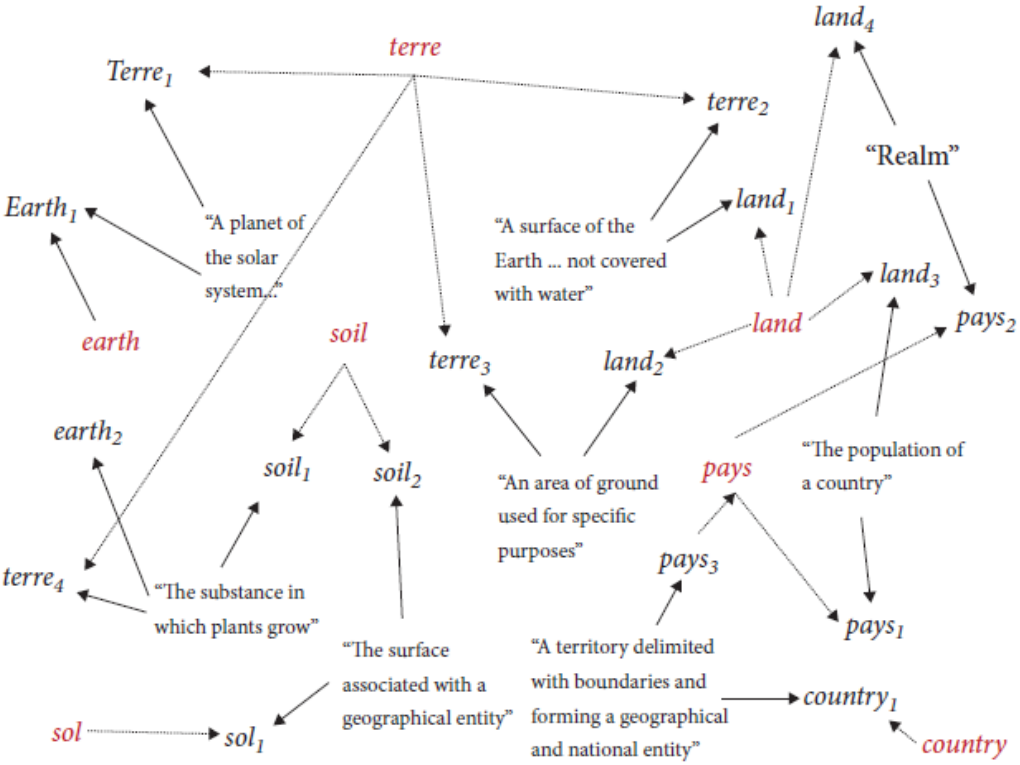
applies to a small object, it translates into French as “*clé*”. However, when it designates the part of a keyboard, it translates as “*touche*”.

ii) A lexical item conveys (at least) two different meanings: these meanings are connected to different fields of knowledge. The noun “dump” is used in waste management and in computing. In waste management, it is defined as ‘a specific place where waste is placed’ and translates into “*dépotoir*” or into “*décharge*” in French. In computing, it designates ‘an operation that consists in emptying the memory’ and its French equivalent would be “*vidage*”.

iii) A lexical item carries (at least) two different meanings: these meanings coexist in the same domain. The example of the French “*terre*” used in the domain of the environment illustrates this situation. “*Terre*” has four different meanings: ‘a planet of the solar system inhabited by living organisms’, ‘a surface of the Earth occupied by continents or islands and not covered with water’, ‘an area of ground used for specific purposes’, and ‘the substance in which plants grow’. The first meaning of “*Terre*” translates into English as “Earth”; the second as “land”, the third also as “land”, and the fourth as “earth” or “soil”.

As one can imagine, cross-linguistic analyses, in these cases, leads to a certain degree of disorganization. Figure 18 was used by L’Homme (2020) to illustrate a situation in which one compares some of the meanings of a set of polysemous words starting with the four meanings of “*terre”* mentioned in iii. Some of these meanings lead to polysemous English equivalents that can be further associated with French polysemous equivalents, which then leads us to infinite possibilities of equivalence. In Figure 18, lexical items are indicated in red, and lexical units are presented with a sense number. There is also a short description of each meaning which is provided between quotation marks. Equivalents can be found when two lexical units in English and in French are connected to the same meaning. The different connections produce a very complex network of equivalence relationships where meaning distinctions hardly overlap in English and French. L’Homme (2020) considers only part of the meanings attached to the lexical items, so the real situation is even more complex.

Figure 18 – Polysemous items and cross-linguistic relationships



Source: L'Homme (2020, p. 234).

As expected, this scenario leads to problems and challenges when establishing equivalence, since such a complex phenomenon becomes even more complex due to the existence of synonymy. León-Araúz & Faber (2014) list ten problems of equivalence when dealing with specialized language. These problems are addressed further on in this chapter, but we would like to begin throwing some light on the matter by eliciting the three main problems listed by L'Homme (2020).

The first one is the problem of non-equivalence, when one language lacks an adequate equivalent to express the meaning conveyed by a term in another language. It can happen for various reasons (from the lack of a lexical unit to designate the concept to the lack of the concept itself due to cultural differences) and it can be handled in different ways. The translator can use an expression to explain the meaning, borrow the word from the source language, adapt the term, or even use an approximate equivalent.

The second problem is the one of partial equivalence, which can happen because the meanings of terms in languages 1 and 2 do not perfectly overlap. A classic example of partial equivalence occurs when a language makes a distinction

that is not made by the other. For example, Portuguese has two distinct terms to refer to “wood”, one is “*madeira*” (designates wood used in construction) and the other one is “*lenha*” (designates wood used to make fire). In English, this distinction does not happen and the term would require an explanation. Partial equivalence always depends on the pair of languages at hand. Using the same example to compare Spanish and Portuguese, we now have a case of exact equivalence because Spanish also makes this distinction by using the terms “*madera*” and “*leña*”. Still according to L’Homme (2020), this example is just one of the many possibilities that partial equivalence can present. In this case, the term meaning in one language is much more general than that of the other language. In fact, the meaning of “wood” in English includes both meanings of “*madeira*” and “*lenha*” in Portuguese, but different cases of partial equivalence sure do exist.

Lastly, there is the issue of structural divergences. They happen when the equivalence candidates belong to different parts of speech, which makes equivalence even more complicated. For instance, “bookmark” can be used both as a verb or a noun in English. However, when you try to translate them to Portuguese, you need different terms to designate the verb and the noun. “Bookmark” as a noun would be translated as “*marca páginas*” and the verb would be translated as “*marcar a página*”. An even more complicated term is “Google”, which in English became a verb (to google something) but in Portuguese still requires the verb “search” to make the activity clear (*pesquisar/buscar no Google*).

L'Homme (2020) also includes the difficulties posed by term variations in equivalence. A term in one language, for example, can have more than one equivalent in another language (which is the case of partial equivalence exemplified previously) and the root of this issue relies on the variation of terms. And because variation occurs in every language, it is also a factor that has to be taken into consideration by translations as it will inevitably challenge the translation work.

This brings us to other issues when it comes to equivalence in translation. One of them is the types of equivalence one could find when building a linguistic resource based on specialized lexicon. We aim to investigate this phenomenon in this scope and to predict some possible equivalence types which may arise in our analyses and what difficulties they can pose considering the conceptual differences between the languages and cultures involved. The first challenge in translation has

been briefly explained: one must pay attention to the context in order to get the correct terms in the target language.

To continue discussing the matter, Espinoza *et al*. (2009) came across what they defined as translation contexts in ontology localization:

i) The existence of an exact equivalent for a lexical unit.

ii) The existence of two or more context-dependent equivalents, which could be regarded as synonymy.

iii) The existence of a conceptualization mismatch when, as exemplified before, the conceptualizations are not shared across languages and cultures.

Although the first situation is what every translator would consider ideal and would hope for, or even imagine to be a common scenario in the specialized translation field, it is quite difficult to find. The second and third types, however, are commonly seen. According to León-Araúz & Faber (2014), the second and third types of translation contexts outlined by Espinoza *et al.* (2009) are interconnected translation problems. The authors list ten problems a translator could find when it comes to cross-lingual meaning, even in specialized domains. Table 5 explains and exemplifies each one of the problems, as highlighted by León-Araúz & Faber (2014).

Accounting for each one of these complex translation problems is a hand full. That is why ontologies are valuable: they help disambiguate lexical units by organizing them in terms of the conceptual and contextual information they provide. As stated by Montiel-Ponsoda *et al.* (2011), when there are multiple options one could choose from in each language, it is necessary to correlate which term in language A is the appropriate translation of the variants in language B in a way that the difference between all of them is clear and unambiguous. This is the part of the translation work which attributes significance to the translation relations between terms. However, as previously mentioned, even when this stage is complete, there is not a guarantee that there will be a perfect, complete equivalence. A relation in which a term in language A has the same sense correspondence as a term in language B, in which A = B and B = A, may still not be possible. As León-Araúz & Faber (2014) exemplify, the translator still has to take the variables in consideration because “if a concept is designated by an informal term variant, it should not always be translated by its informal counterpart in another language and vice versa because one must

Table 5 – Problems in cross-lingual senses regarding concept & term dynamics

Translation problem	**Example**
The entity exists in both cultures, but the term for it in one language culture is more general or more specific than the other.	"Shingle" in English is a term that covers several more specific terms in Spanish.
The entity exists in both cultures, but only one language culture has a term for it.	"River" and the French "*fleuve*" and "*rivière*" and "*espigón*" and "*jetty*" and "*groyne*".
The entity exists in both cultures, yet the terms are not exact correspondents because they highlight different aspects of the concept or focus on it from different perspectives.	The French "*fleuve*" and the English "main stem".
The entity exists in both cultures and both language cultures have terms for it, but only in one language the concept has been lexicalized in several variants with different communicative or conceptual consequences.	The Spanish "*intestinos*" and the English synonyms "intestines" and "bowels" or "rubble-mound breakwater" and the Spanish synonyms "*dique de escollera*" and "*dique en talud*".
The entity exists in both cultures and both language cultures have terms for it, which approximately correspond. However, the lexical categories appear to have different structures in each culture and thus seem to operate on different design principles.	"Dock", "quay" and "wharf", and the Spanish "*muelle*", "*embarcadero*", and "*dársena*".
The entity exists in both cultures, but its cultural role (utility, affordances, and hindrances) in each one is different. This leads to a conceptual mismatch and lack of correspondence.	"*Pier*" and "*embarcadero*".
The entity exists in only one of the cultures, but its name has been adopted in the other culture to refer only to the foreign culture-specific concept.	The Australian "billabong", the African "dambo" or the Canadian "muskeg".
The entity exists in both cultures, but one culture has recycled a term from the other culture to refer to another totally different concept.	"*Playa*" in West US as "dry lake" and not as the usual Spanish equivalent "beach", but "*salar*".
The entity exists in only one of the cultures and is totally unknown in the other without any designation.	"*Pejerrey*", a fish that only can be found in South America.
The entity exists in both cultures, but one of the cultures may refer to it with a metonymic designation and be ambiguous.	"*Groyne*" as the equivalent of the Spanish "*escollera*", the material it is usually made of.

Source: León-Araúz & Faber (2014, p. 11).

also consider the nature of the communicative situation." (LEÓN-ARAÚZ & FABER, 2014, p. 12).

Cimiano *et al.* (2010) also expressed their concerns regarding this matter and stated that unintended changes in meaning may occur if the term that was elected as an equivalent has different connotations in the target language. Since multidimensionality has an enormous impact on how concepts can be classified and also on how variants emerge, it can end up impairing equivalence, as explained by León-Araúz & Faber (2014). Both these authors work in the Multilingual Semantic Web project and have used the translation relations proposed by Montiel-Ponsoda *et al.* (2011) – descriptive translation and cultural translation – as basis for a more extensive classification and description of translation relations. They elected nine different types and described them as follows, taking into consideration the problems previously explained by table 2:

1. Canonical translations apply when no equivalence problems arise, and the translation relation may be symmetric. "River" and "*río*" would be canonical symmetric equivalents, but this does not mean that when canonical translations are found no other relations are possible, since context can impair the degree of equivalence.
2. Generic-specific translations would address problems 1, 2 and 3 which are related to cross-lingual categorization differences, depending on the communicative situation and directionality. A specific-generic translation would apply when translating the term "shingle", which in Spanish can be translated by its hypernym "*material de grano grueso*" (coarse material). In the same way, when the context describes a beach nourishment scenario in Spanish, the term "*material de grano grueso"* can be translated by its canonical form "coarse material", but also by its specific translation "shingle". Alternatively, the following relation may apply.
3. Extensional translation would address problems 1 and 2 and is a kind of generic-specific translation, because the original term is translated by all of the hyponyms of the concept in the target culture. In this way, "shingle" can also be translated by the enumeration of its subtypes (*arena y grava*).
4. Communicative translations would address problem 4 establishing register correspondence among domain-specific and diaphasic variants. The canonical translation of "*lodo*" is usually "mud", but in a water treatment domain, experts have a preferred designation: sludge. Furthermore, depending on the communicative

situation, certain terms can be translated as the expert neutral variant or the lay-user variant in the target language (e.g. “intestines” or “bowels” for “*intestinos*”).

5. Functional translations would address problems 5, 6 and 7 and involve deculturalising original terms so that receivers can relate to the concept. “Muskeg” can be translated as “*turbera*” and “*malecón*” as “seawall”. These equivalents lose their cultural traits but are the closest concepts in target cultures from a semantic point of view. Other terms, such as “quay”, “dock” and “wharf” must rely on additional contextual features, since they can all be translated as “*muelle*”, “*embarcadero*” and/or “*dársena*” depending on the size, function and position of the structures. This relation is particularly asymmetric. For instance, “*turbera*” could hardly ever be translated as “muskeg”, since unless the communicative situation points to this particular type of Canadian wetland, the canonical translation “bog” would apply in most of the cases.

6. Cultural translations apply when cross-cultural differences impair the translation process and affect both concepts and terms. They would be another way of addressing problems 6, 7 and 8 and consists of adapting original culture-bound terms to other culture-bound terms in the target culture. The usual canonical translation of “pier” is “*embarcadero*”, but piers are often recreational areas that do not fit with the Spanish concept. In these cases, the most suitable translation would be “*paseo marítimo*” (literally boardwalk), or even “*malecón*” or “*costanera*” for South American Spanish, since even if these kinds of constructions are slightly different, the cultural component of the concept is preserved.

7. Descriptive translations would also address culture-bound problems and make explicit certain semantic features according to user communication needs (problems 7 and 8) or in order to distinguish a concept that has not been termed in the target culture (problem 2, 9). For lay users, the term “muskeg” could be translated as “*humedal canadiense muskeg*” (Canadian wetland muskeg), adding and highlighting its hypernym and location. In contrast, the term “jetty” can be translated as “*espigón*”, which is the canonical translation of “*groyne*” or even “*dique*”, which would be a functional translation according to its general nature and the wide array of functions it may have. However, if both terms are found in a text (“jetty” and “*groyne*”), some distinction must be made. In this sense, a descriptive translation could be “*espigón de encauzamiento*”, which explains the particular function of “jetties”.

8. Non-translations also address culture-bound problems (7, 9) when entities and/or lexicalizations do not exist in the target culture (*pejerrey*), but also in specialized communication. Terms like “muskeg” or “billabong” can be kept in their original form if the receivers are experts who do not need any description or contextualization.
9. Metonymic translations would address problem 10 and apply when original terms are expressed in the form of a metonymic variant and target terms are not. “*Groyne*” could be translated both as “*espigón*” or “*escollera*” (metonymic variant), but “*escollera*”, in its coastal structure sense, can only be translated as “*groyne*”.

It is important to keep in mind that, as previously emphasized in this subsection, even when translations relations are highlighted and analyzed, there may be situations where new translation difficulties could arise, depending on numerous factors, including the material being translated. However, we believe León-Araúz & Faber (2014) were successful in enumerating the most common relations, precisely because they acknowledge the complexities surrounding equivalence. They do not ignore the dynamism of translation pairs, on the contrary: the authors recognize that complete and integral symmetry across languages is hardly ever achievable. As pointed out by Gangemi (2012, p. 1), “[…] if we envisage applications that are cross-linguistic, they need to work at the level of cognitive relevance, not at that of single, decontextualized data or term equivalences”.

These characteristics of lexical equivalence and translation problems aimed at Terminology combined with the semantic relations explained by us and touched on by León-Araúz & Faber (2014) are going to be considered in our bilingual analysis. Thus, considering the contributions mentioned so far, we believe that we can begin our analysis of term alignment and lexical substitution with the data provided. The next chapter will be dedicated to explaining how the analyses will take place taking into account the points described in this segment of our study.

4 METHODOLOGY

Referring back to the purposes of our research, our primary objective is to investigate the phenomenon of lexical variation in both Portuguese and English when it comes to the term alignment and lexical substitution stage in NLP. Among the secondary objectives there is an intended analysis of what types of lexical variation occur during the alignment and substitution processes and the aim of finding out to what extent the multilingual aspect, especially lexical equivalence, is affected by said lexical variants. Consequently, after having described the theoretical background behind this Computational – Linguistic interface, this chapter addresses the methodology used by us in order to accomplish our objectives. Chapter four aims to outline how the analyses of lexical variation and lexical equivalence will take place and what materials are going to be used to accomplish these analyses.

In order to delineate our methodological choices, we will begin by describing Corpus Linguistics and its usability, since it is one of the sources of linguistic information we used in this stage of the study. Additionally, we will describe the characteristics of our corpora and how the tool selected by us – Sketch Engine[1] – was used in this stage of *corpora* analysis. This information will compose the first part of our chapter, numbered section 4.1. In the second half of our chapter, numbered section 4.2, we will list the selected terms used in this study, both in Portuguese and in English, and outline the reasons behind our terminological choices. We intend to also detail the computational experiments applied to these terms and their stages. Lastly, we will detail the analytical steps employed for our linguistic analysis and classification of the terms suggested by the models used in this study.

4.1 Materials: Corpus Linguistics & Sketch Engine

Corpus Linguistics can be described as a methodological approach[2] - widespread in Linguistics studies - which involves computer-based empirical analyses of language by employing large and electronically collections of text, called corpus. The texts that compose a corpus are examples of natural language in written

[1] https://www.sketchengine.eu/. Access on 11/08/22.

[2] Corpus Linguistics is perceived as both a methodology and a theory. The first approach, adopted in this work, is known as corpus-based and it uses corpora to analyze, test, and improve linguistic theories defined early. The second one, known as corpus-driven, lets the corpus dictate what is going to be analyzed based on the information that appears.

context and can have different sizes and content depending on the objectives behind its compilation. Berber Sardinha, who is the scholar responsible for introducing Corpus Linguistics in Brazil, describes a corpus as "[...] a set of textual linguistic data that have been carefully collected with the purpose of serving for the research of a language or language variation." (BERBER SARDINHA, 2000b, p. 235, our translation)[3]. Li (2015, p. 465), describes a corpus in similar ways. According to the author, it can also be understood as "[…] a large collection of machine-readable texts compiled with a specific purpose that can be retrieved with particular computer software for linguistic research".

We believe that both definitions are complementary for they bring up relevant information about the corpus compilation stage. Both Berber Sardinha (2000b) and Li (2015) emphasize the goals that the compiler has in mind. Consequently, the complied corpus will be structured, organized, and stored according to its purpose. The selected text excerpts will be included in the corpus considering their usability. If one aims to investigate and/or compare the differences between the passive voice in Portuguese and the passive voice in English, for example, it would be necessary to retrieve texts in which transitive verbs occur in high frequency and texts which display formal writing, for example. Otherwise, the passive voice would not be adequately shown in the corpus, and it would not be a satisfactory source of linguistic information regarding this writing style.

Li (2015) also mentions the relevance of machine-readable texts since computers play an important role in the storage and analysis of corpora. The software used by us will be explained later, along with the types of linguistic analysis one may execute using Corpus Linguistics as a methodology. The main reason why we chose Corpus Linguistics as a methodological approach is its relevance to Translation Studies. The field turned to corpora use in the 90s with the interest shown by Mona Baker (1993), who began the corpus-based translation studies. One of the reasons why Translation Studies turned its attention to corpus usage is the fact that cross-lingual analyses can largely benefit from corpora usage due to the amount of linguistic information present in bilingual corpora. This information can be easily extracted by the software being used in the analysis and a well-compiled corpus can be a rich source of translation aid.

[3] Originally: "[...] conjunto de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística."

Since our goal is to analyze the quality of term alignment and lexical substitution in monolingual and bilingual tasks, we must describe the three main types of corpora used in Translation Studies. Due to the fact that it is a recent area of study, terminological differences can be spotted when referring to these corpus types. According to Granger (2003), the differences between corpora nomination happen because Translation Studies and Contrastive Linguistics both use these denominations. Since our study is partially focused on translation, we will identify the types of corpora used in this field of study.

The first one is called translation corpus, also known as translational corpus. According to Baker (1999) it refers to a corpus of translated texts. This data would convey the same semantic information and would work as a resource for establishing equivalence and equity of terminology between two or more languages. This type of corpora can be easily aligned and then compared. The problem with translation corpora is that they are difficult to find, and the genre is limited to official documents and very few literary works. In countries such as Hong Kong and Canada, one can find the governmental documents translated to all of the official languages. This, however, does not happen to social media posts, news, articles, and other genres that one may want to use to compile a corpus. This is the main limitation of translation corpora. Figure 19 exemplifies this type of corpus. This example comes from one of the corpora compiled by us and, as we can observe, the texts are written in a way that they express the exact same message, but in different languages. The limitation relies on the fact that these documents were created with this intent, and other translated texts such as this one are not easily found.

The second type of corpus is known as comparable corpus, and it is composed of original texts in the different languages. These documents, however, are not translations of one another, simply texts that refer to the same subject, belong to the same genre, and can be found in similar online environments. Despite the fact that alignment in this case is impossible, these texts pose the advantage of not being influenced by other linguistic systems. The main challenged offered by comparable corpora is the difficulty in establishing equivalence. This type of corpus is frequently confused with translation corpora. Yet, researchers seem to be close to reaching an agreement on how to distinguish one from the other. According to Li (2015, p. 471), comparable corpora can be described as “[…] thematically parallel

non-translational corpora". Figure 20 exemplifies this type of corpus and, once again, the example comes from one of our corpora.

Figure 19 – Example of translation corpus

PRIVACY DIRECTIVE – BRADESCO ORGANIZATION

Purpose:

The Privacy Directive of Bradesco Organization, hereinafter referred to as "Organization", was created to demonstrate its commitment to the protection of your personal data and privacy. We will show below how your information will be collected, used, protected, and what your rights are and how they can be exercised.

DIRETIVA DE PRIVACIDADE – ORGANIZAÇÃO BRADESCO

Objetivo:

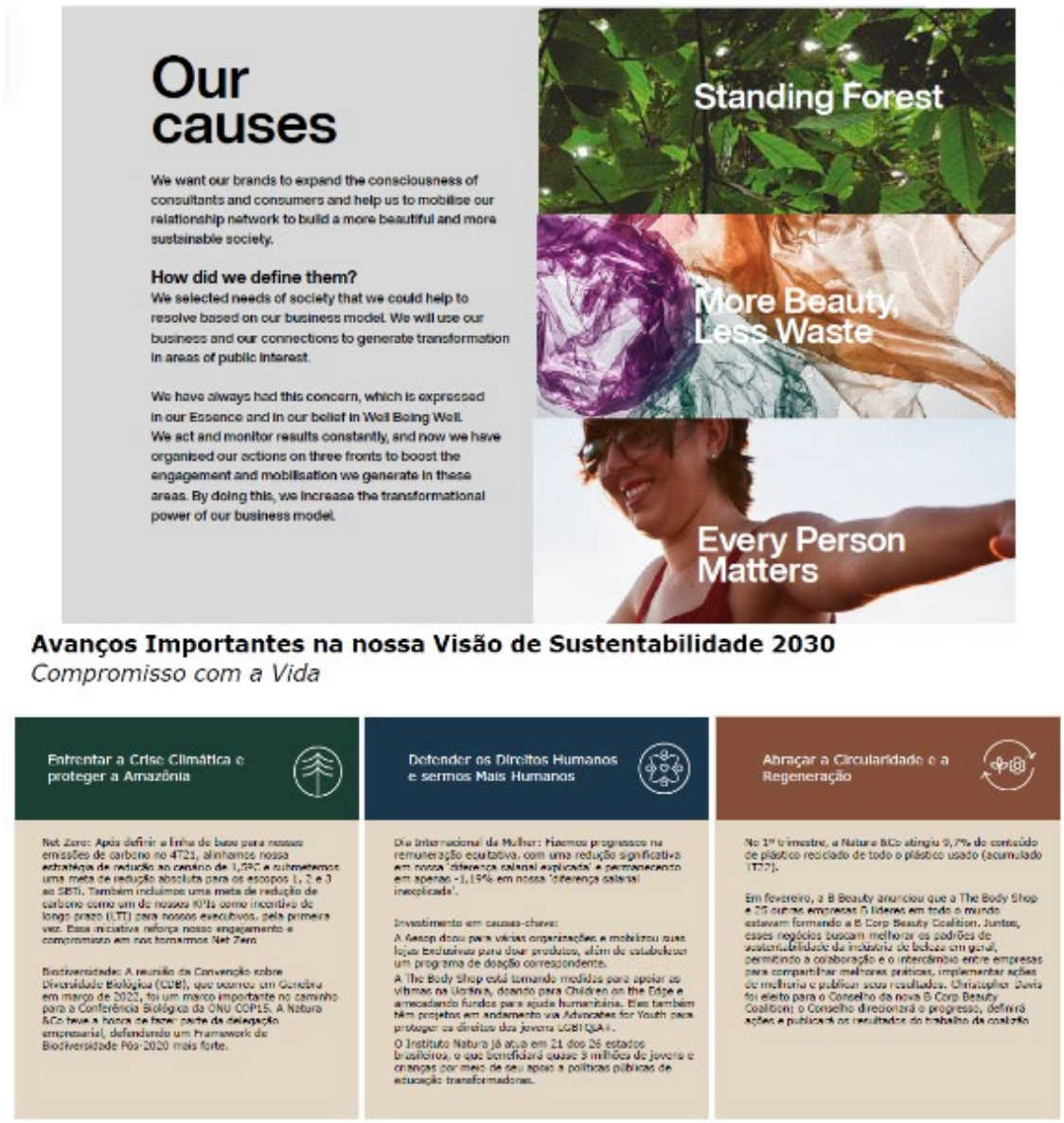
A Diretiva de Privacidade da Organização Bradesco, doravante chamada de Organização, foi criada com o intuito de demonstrar o seu compromisso com a proteção de seus dados pessoais e privacidade. A seguir apresentaremos como as suas informações serão coletadas, utilizadas, protegidas, bem como quais são os seus direitos e como eles poderão ser exercidos.

Source: Bradesco official website (2022).

In this case, we provide an example of texts which are not identical in terms of text, in the sense that they are not translations of one another. However, they display the same type of content. Lastly, we must mention parallel corpus. This type of corpus seems to be the source of bigger and more considerable confusion, since the term has been used to describe different types of cross-lingual corpora. We align with Li (2015, p. 473), who describes parallel corpora as "[…] any collection of texts in different languages and language varieties conveying similar information produced under similar pragmatic conditions". As stated by the author, parallel corpora can be used as an umbrella term for cross-lingual corpora, due to the fact that it can include translation corpora and samples of the same genre for different languages, for example. Therefore, parallel corpus can include both a translation corpus, which is translated, passible of alignment, and contains identical texts, and a comparable corpus, which is translated, not passible of alignment, and contains similar texts.

Thus, both the documents showed in Figures 19 and 20 could be part of a parallel corpus.

Figure 20 – Comparable corpus



Source: Natura official website (2022).

Having explained the corpus types one can work with when using Corpus Linguistics aimed at Translation Studies, we now describe the differences between quantitative and qualitative research in corpus-based translation studies. Quantitative methods, as defined in Li (2015, p. 474),

> […] are preceded by the researcher's ideas and hypotheses about observed dimensions to calculable and measurable parameters. Frequency occurrence of a language form, its combinations with other items in

> discourse as well as patterns of semantic similarity, oppositeness and inclusion all contribute to a language-specific character of SL and TL forms.

Statistical analysis, for example, is considered a quantitative method of analysis. Lewandowska-Tomaszczyk (2011) lists seven translational quantitative criteria used on their work: lexical unit frequency, keyness, frequencies of syntactic patterns, frequencies of classes of lexical-semantic patterns, frequencies of types of figurative extensions, quantitative cross-correspondence of concepts from the same conceptual cluster, and distributional criteria. Qualitative study, on the other hand, "[…] is based on interpretations of resemblance between concepts presented in the original SL and TL translation from the experiences, actions and observations of individuals." (LI, 2015, p. 475). Examples of qualitative analysis would be the five selections studied by Sinclair (2004), which are core, collocation, colligation, semantic prosody, and semantic preference.

The difference between quantitative and qualitative analysis, according to Li (2015, p. 475), is that "while quantitative research investigates relations between a few variables in larger samples, qualitative research deals with relations between many variables that can be investigated in smaller samples." Additionally, it can be said that while quantitative analysis looks at the numbers and the information provided by them, qualitative analysis is concerned with information that does not revolve around quantity and/or frequency.

The choice between quantitative and qualitative methods relies on the goals and purposes of the research being conducted. Each method has its advantages and disadvantages and can contribute to the study in different ways. Moreover, it is possible to combine both methods and look at the information in the corpus both in a quantitative way and a qualitative one. Usually, this combined approach is what commonly happens in corpus-based translation studies.

In summary, Granger (2003) lists the main benefits brought by corpora use in Translation Studies as it being a good source of information, a good source for equivalence, terminology, and phraseology, a source of large quantity and considerable coverage of genres and texts, and an easily usable tool for retrieving linguistic and contextual information. These are the main reasons why we opted for using Corpus Linguistics as a methodology in this work, since we agree with the definitions of Granger (2003).

This brings us to the corpora used in our study. The chosen scope of this work is the retail industry, because we have been working with this domain in the VLHSem Project, and the terminology to be analyzed is part of this domain. It is the scope appointed to us by the company and we have decided to commit to it. We opted for two different corpora that combine online, free documents found on national and international companies' websites. These documents were made available for the public by the companies; thus, we could gather them at the official websites.

We attempted to select documents that would contribute to the terminology study we are conducting. One of the corpora, named "Corpus_Retail_English" is composed of documents related to the retail industry and all of the texts are in English. The second corpus, named "Corpus_Retail_Portuguese" is also made of documents related to the retail industry. This corpus, however, is formed completely by documents in Portuguese. Table 6 shows the relevant information about the texts selected to compose our corpora.

Table 6 – Corpora characteristics

Characteristics	**Corpus_Retail_English**	**Corpus_Retail_Portuguese**
Language	English	Portuguese
Genre	Annual reports, press releases, financial reports, transcripts of conferences and meetings, analyses of results	
Number of words	927.336	509.445
Tokens	1.234.809	661.087
Number of documents	22	21
Companies	3M, Johnson & Johnson, Nestlé, Nike, Carrefour, Adidas, BMW, Harrods, Colgate/Palmolive, Liverpool Mex, Coca-cola, Mercedes-Benz, BASF, Boticário, Burger King	Nestlé, Carrefour, BASF, Boticário, Burger King, Natura

Source: made by the author (2023).

As mentioned previously, the documents were retrieved from the companies' official websites. Generally, the documents are located in the “Investor” tab, which is a way that the companies have of presenting its profits and products to potential and future investors. All documents used to compile our corpora belong to public domain, since the companies themselves release the information on their website. Therefore, the documentation does not violate ethical regulations and policies.

These documents are formed by data that provides input of the financial aspects of the companies, product information, marketing strategies, and profit evidence. This documentation is aimed at shareholders who might be interested in financially investing in the companies and are usually robust, since the information has to the complete and proof of financial gain has to be shown. We believe that the intended access to stockholders and investors must be one of the reasons why this documentation is freely available online.

It is also worth stating that some of the selected companies provide this documentation in more than one language, usually both Portuguese in English. It is the case of Boticário, Nestlé, Carrefour, and Natura, to name a few. Therefore, in our analysis, we used both translation corpus and comparable corpus, according to the denomination used by Li (2015). The chosen companies were selected considering their importance in the national and international market, as well as the usability of the material provided in the websites.

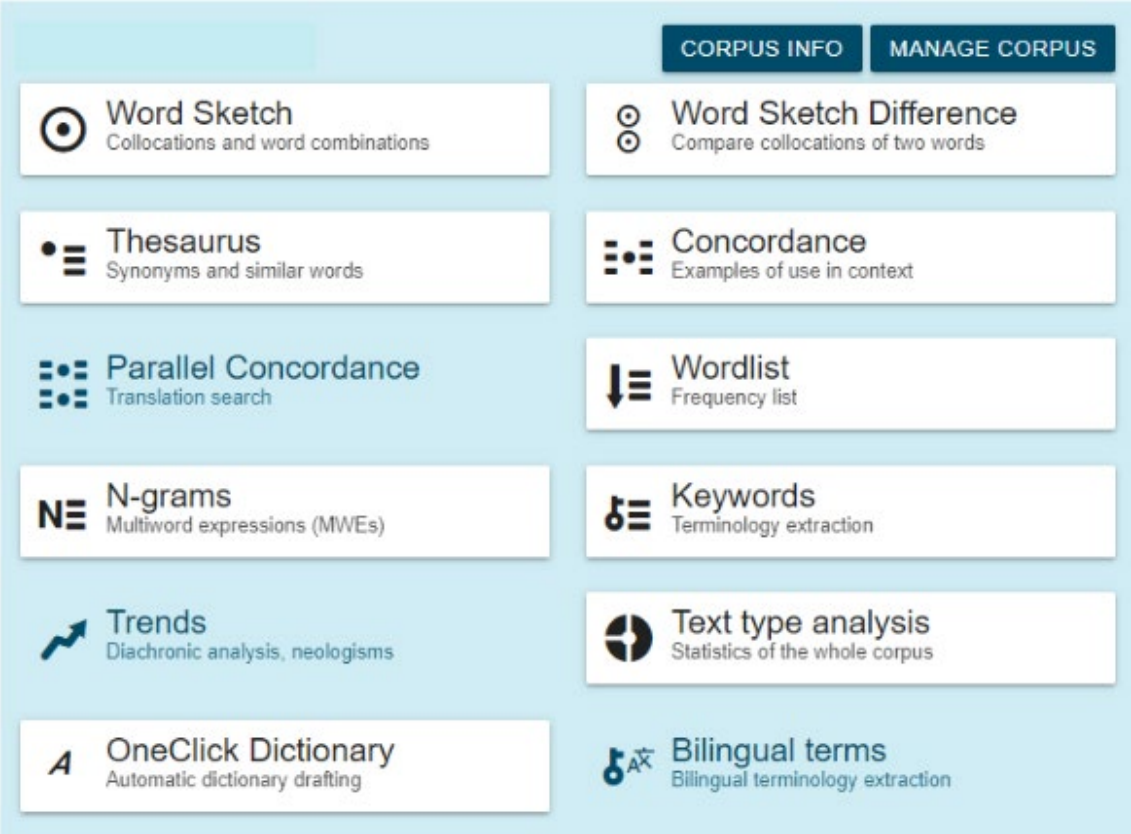
The apparatus used by us in order to work with the corpora is called Sketch Engine[4]. Sketch Engine is a tool developed by a research company called Lexical Computing. Sketch Engine is described in the Lexical Computing website as a leading corpus management and corpus query tool used by many linguists, lexicographers, translators, and publishers worldwide. The reason why it is so popular among researchers is due to its functionalities together with the scalability, multilingual support, and ability to handle the largest available corpora.[5] Additionally, Sketch Engine contains 600 ready-to-use corpora in more than 90 languages, each having a size of up to 60 billion words to provide a truly representative sample of

[4] This tool was chosen by us due to its completeness when it comes to the possibilities of corpora analysis. This choice was made by VLHSem Group and we decided to extend its usage to the scope of this work.

[5] Information retrieved and adapted from: https://www.lexicalcomputing.com/lexical-computing/. Access on 09/11/22.

language.[6] Portuguese and English are among the languages represented in the tool. Figure 21 shows the features of Sketch Engine.

Figure 21 – Features of Sketch Engine.



Source: Sketch Engine (2022).

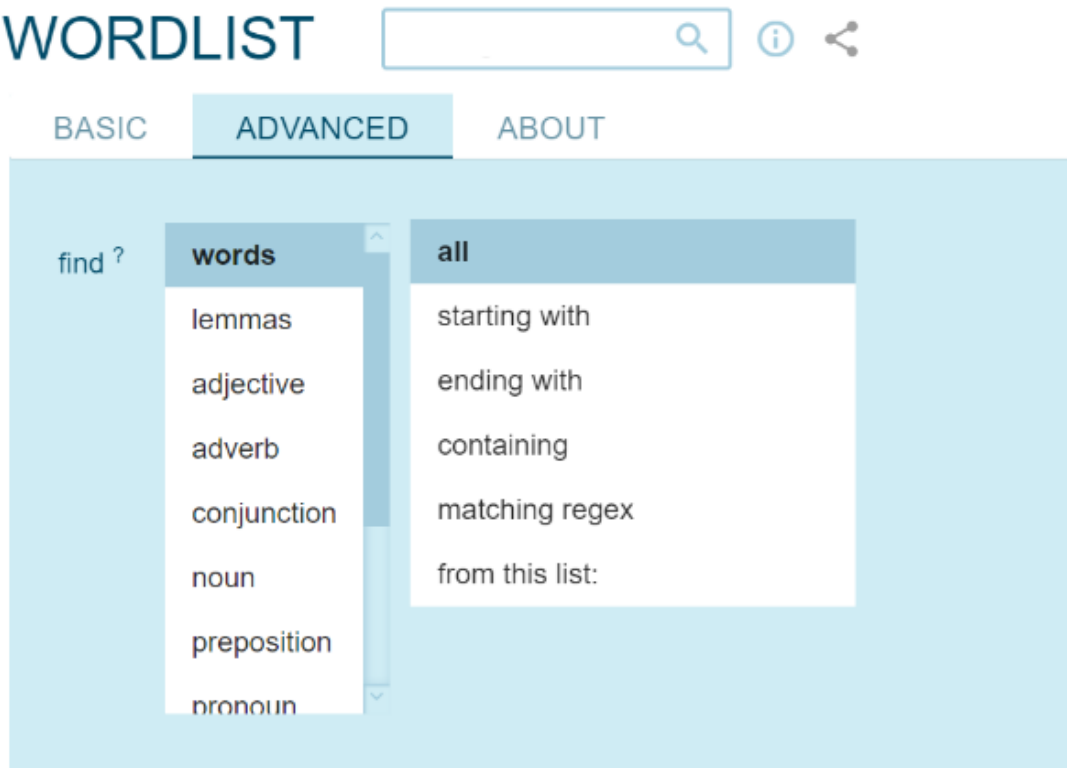
As we can observe, there are multiple features which make Sketch Engine a valuable tool for our analysis. We will describe with more detail the features which are relevant for our work, and which are going to be used by us. The features which we will not use are not going to be detailed. This does not mean, however, that they are somehow irrelevant or useless.

The first feature we would like to discuss is the “Wordlist”. It is a list which displays the frequency of occurrence of each word in order, and which can be used in a quantitative approach to corpus analysis. It is important, however, to eliminate some of the word categories that could “pollute” the list, such as prepositions and articles, if they are not the subjects under linguistic investigation. The reason why is because terms such as “an”, “a”, “the”, “of”, and so on are usually the ones that happen more frequently in every text. Figures 22 and 23 show how you can begin

[6] Information retrieved and adapted from: https://www.sketchengine.eu/#blue. Access on 09/11/22.

your search using the “Wordlist” feature and what the results in the “Corpus_Retail_Enlglish” were. These results were not filtrated before being displayed for the purpose of exemplifying what happens when some categories are not eliminated from the search.

Figure 22 – “Wordlist” set up



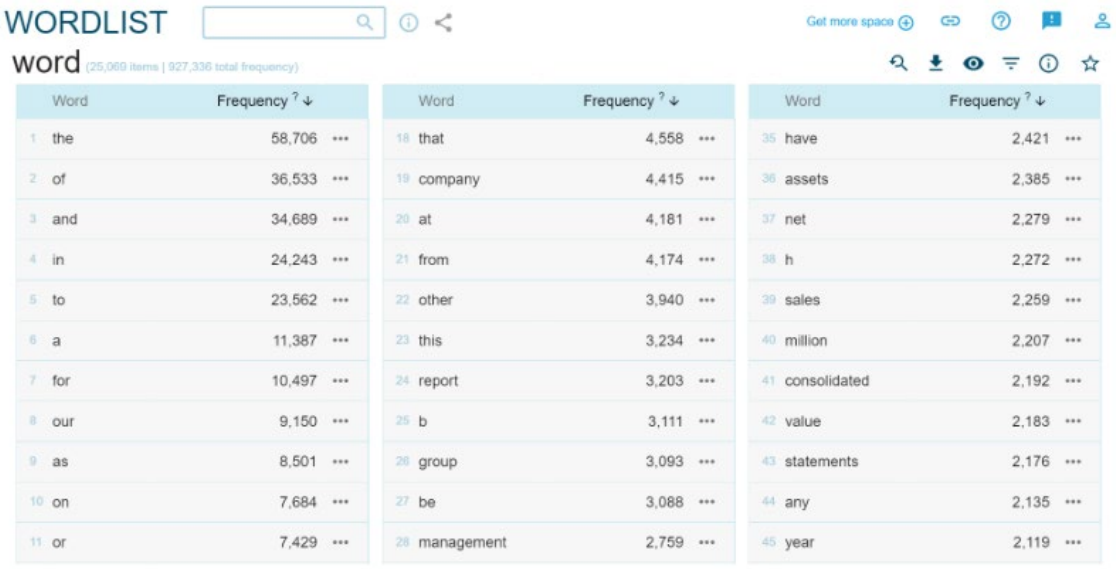
Source: Sketch Engine (2022).

The “Wordlist” feature can be useful in terminology studies because by using this feature, linguists have a chance of sorting out the most common (and uncommon) terms used in a certain domain or culture. In order to successfully accomplish this, the compiled corpus must be rich, vast, and relatively complete in terms of content, so the terminology is successfully represented.

The second type of feature worth mentioning is “Word Sketch Difference” (WSD). This specific feature is extremely useful in our work because it is fruitful for works regarding semantic analysis since it makes it easy to spot close synonyms, antonyms, and words belonging to the same semantic field. WSD generates word sketches for two chosen words and compares them making it easier to observe differences in use. The result assigns red and green shades to each lemma. The collocates in green tend to combine with the green lemma while the collocates in red

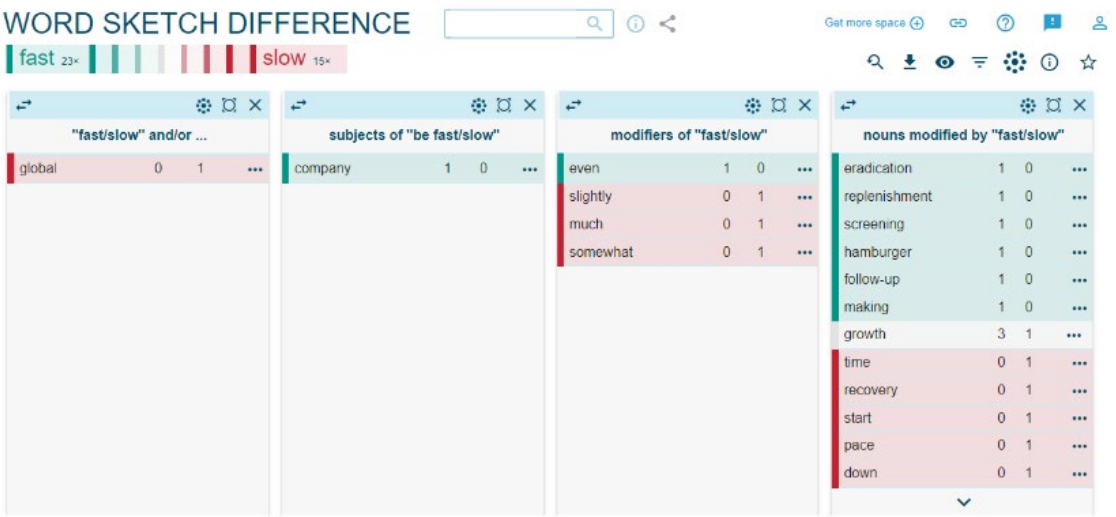
tend to combine with the red lemma. The white collocates combine with both lemmas. Bolder shades of green and red indicate stronger collocations. Figure 24 shows how the results are displayed. We searched for the pair “fast” & “slow” which can be considered antonyms.

Figure 23 – “Wordlist” results



Source: Sketch Engine (2022).

Figure 24 – Word Sketch Difference results for “fast” & “slow”

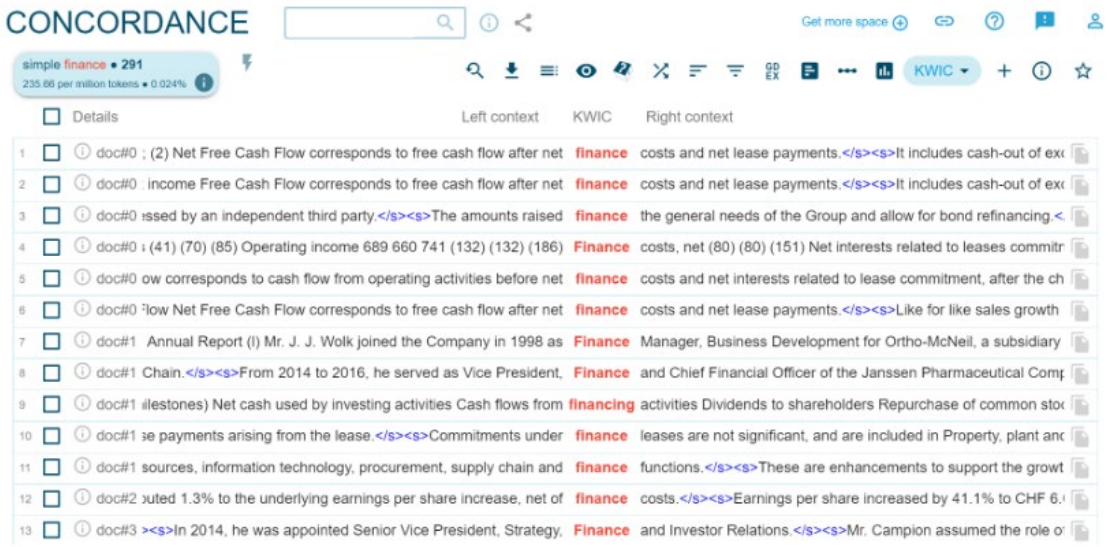


Source: Sketch Engine (2022).

Lastly, we would like to introduce the usability of “Concordance”, also a significant feature for our analysis. It can be used to find examples of a word, lemma, phrase, tag or even a complex grammatical or lexical structure. It is possible to see the words surrounding the searched item and its context by clicking in the term in red. The basic search allows one to look for a simple word or phrase while the advanced tab offers more detailed options for setting search criteria, such as query types, subcorpus, macro, filter context, and text types. It is also possible to see the number of times the word appears on the corpus. Figure 25 exemplifies results for the term “finance” in our Corpus_Retail_English search. As it can be observed, “finance” is a term that occurs 291 times.

Its usability relies on the investigation of syntactic and morphologic patterns, related terminology, surrounding words, frequency lists, among other traits of the terminology. In summary, it allows the linguist to look at the “behavior” of the terminology, its characteristics, and its placement in a sentence or in the larger portions of text.

Figure 25 – Concordance results for finance



Source: Sketch Engine (2022).

As one can conclude, the tool chosen by us is rich in terms of features and opportunities of linguistic analysis. The features we will use the most were briefly outlined here for the purpose of completing the linguistic analyses we are compromised with. Moreover, Corpus Linguistics offers an abundant scope of linguistic analysis considering the extent of our purposes. In the next section, we will

detail the experiments made by three of our models and the main elements of the methodological steps applied in our analysis.

4.2 Methodological steps: terms & semantic analysis

As previously stated, this section aims to outline our methodological steps additionally to describing the executed experiment with three of our models and our strategy when selecting the terms to be analyzed. Thus, we will begin by describing the terminology used and the models. Lastly, we will approach the analysis organization.

The terminology selected by us belongs to the retail domain, as previously stated. Our choice was made taking into consideration factors such as the material available for analysis, knowledge of the domain, and knowledge of the related terminology. We selected two pairs of terms from a list of lexical substitution given to the VLHSem Project team and these pairs are “plant – site” and “material – article”. We have also outlined their possible synonyms and their possible translations to Portuguese. They are displayed in Table 7. It is worth noting that the possible synonyms and possible equivalents were delineated by us based on our knowledge of the language and knowledge of the retail domain. The actual suggestions provided by the models will be displayed later, but since our models operate exclusively in English and do not provide equivalents for the terminology in other languages, we have listed possible equivalents ourselves.

As one can observe, two pairs of aligned terms offer a variety of possibilities of synonyms and lexical equivalents even when they are part of a specialized domain. From now on, we must refer to the four terms to be analyzed (material, article, plant, site) as target terms, since this is how the models classify them. Our intent is to avoid confusion regarding the terminology we are referring to. These target terms were studied by us and allocated in their respective contexts when it comes to retail industry. Then, they were used by our models as a source of synonym prediction.

Table 7 – Terminology selected for analysis

Original term	Alignment choice	Possible synonyms	Possible equivalentes
Material	Article	Stuff, product, thing, item, object, unit, commodity, artifact	Material, artigo, produto, coisa, item, objeto, unidade, mercadoria, artefato, bem intermediário, bem de comparação, bem de conveniência, commodity, estoque, granel
Plant	Site	Factory, unit, works, foundry, mill, shop, yard, industrial unit, business unit, building	Planta, site, fábrica, unidade, fundição, departamento, loja, loja física, loja âncora, varejo, unidade industrial, unidade de negócio, prédio, canal de distribuição, fonte de suprimentos, franquia, estoque

Source: made by the author (2023).

There was a variety of lexical substitution models developed by us, but we decided to analyze three of them, considering a basic model, an intermediate one, and a sophisticated model to measure how the addition of linguistic information helped the models. The lexical substitution task is related to the term alignment task, as explained in chapter two. In Computing, this task is concerned with finding appropriate substitutes for a target word in a given context, according to Arefyev et al. (2020, p. 1243). The required solution for this problem is

> […] finding words that are both appropriate in a given context and related to the target word in some sense (which may vary depending on the application of generated substitutes). To achieve this, unsupervised substitution models heavily rely on distributional similarity models of words (DSMs) and language models (LMs). […] It learns word embeddings and context embeddings to be similar when they tend to occur together, resulting in similar embeddings for distributionally similar words. Contexts are either nearby words or syntactically related words".

The three models executed by us follow these premises in accordance with what was stated in chapter two. The models used are called Roberta (a simpler model based on lexical synonyms), Roberta embedding (which could be considered an intermediate model that considers synonymy and word embeddings), and Roberta semantic frames target embedding (one of the most sophisticated models developed by us since it takes semantic frames into consideration as well). The information we provided to the models consists of the target term, the context sentence, and the semantic frame, in the case of Roberta semantic frames target embedding. Then, the models executed the task using lexical synonymy knowledge, word embeddings, and semantic frames. The predictions (the results which consist of the terms appointed by the models and lexical substitutes of the target terms) are evaluated by us and classified as valid or invalid. The criteria used was based on the semantic relations mentioned in chapter three and will be described towards the end of this chapter. We chose these models due to their satisfactory performances according to the literature. For more information about how these models work, see Arefyev et al. (2020).

The predictions presented by our experiment with the three models are displayed in Figures 26, 27, and 28. All these results and our evaluation and classification will be detailed in our analysis chapter further on and these figures represent the totality of data to be analyzed by us. Thus, these results are included here because they represent the terms to be analyzed. The figures outline the target terms, the context, the semantic frames (in cases in which it is necessary), and the predictions.

The methodology used by us to analyze the semantic information of the terms and the accurateness of the suggested synonyms consists of the following steps, applied to each word sense:

i) Compile definitions with the help of dictionaries, thesaurus, glossaries and other lexicographic resources all of the possible senses of each term.

ii) Select the most accurate sense for each prediction considering the context at hand.

iii) Pair the target term and the predictions according to closeness in meaning. A prediction is classified as an appropriate synonym, a hyponym, or a meronym according to the similarities of senses and semantic information in the retail context.

Figure 26 – Roberta

Target	Context	Predictions
Plant	The 140,000-square-foot **plant** is a testament to advanced manufacturing technology.	facility, factory, site, project, production, mill, process, complex
Site	The inauguration of the **site** in Cajamar, the biggest and most advanced integrated cosmetics research, production and distribution center in South America, was a huge success.	project, facility, plant, center, campus, complex, factory
Materials	The solutions for innovation in **materials**, commercial and logistics models, as well as cross-cutting proposals involving more than one topic, were analyzed.	manufacturing, metals, logistics, plastics, aerospace, communications, fuels, construction
Article	The company's newest **article** is selling well, according to the reports we received.	stock, product, IPO, offering, book, issue, story, ad, business

Source: made by the author (2023).

Figure 27 – Roberta embedding

Target	Context	Predictions
Plant	The 140,000-square-foot **plant** is a testament to advanced manufacturing technology.	factory, facility, mill, project, site, production, station, process, planting
Site	The inauguration of the **site** in Cajamar, the biggest and most advanced integrated cosmetics research, production and distribution center in South America, was a huge success.	project, facility, website, location, campus, plant, platform, center, station
Materials	The solutions for innovation in **materials**, commercial and logistics models, as well as cross-cutting proposals involving more than one topic, were analyzed.	metal, plastic, manufacture, logistic, communication, resource, component, fuel, fabric
Article	The company's newest **article** is selling well, according to the reports we received.	story, item, book, product, piece, report, essay, publication, paper

Source: made by the author (2023).

Figure 28 – Roberta semantic frames target embedding

Target	Context	Semantic Frame	Predictions
Plant	The 140,000-square-foot **plant** is a testament to advanced manufacturing technology.	*locale_by_event_scenario*	facility, factory, site, project, mill, process, production, manufacturer, station
Site	The inauguration of the **site** in Cajamar, the biggest and most advanced integrated cosmetics research, production and distribution center in South America, was a huge success.	*locale_by_event_scenario*	location, facility, project, website, locale
Materials	The solutions for innovation in **materials**, commercial and logistics models, as well as cross-cutting proposals involving more than one topic, were analyzed.	*manufacturing_scenario*	metal, manufacture, logistic, plastic, resource, fabric, communication, component, product
Article	The company's newest **article** is selling well, according to the reports we received.	*manufacturing_scenario*	product, item, story, piece, book, publication, report, paper, issue

Source: made by the author (2023).

iv) Evaluate how accurate the predictions were based on the terminological information they convey.

Once we have addressed the model performance, we intend to classify each prediction according to the variant types defined by León-Araúz & Faber (2014) and the semantic relations listed by Murphy (2003, 2010), which are synonymy, hyponymy, and meronymy. These variant types are orthographic variants, diatopic variants (which include orthographic variants, dialectal variants, and culture-specific variants), short form variants (which can be abbreviations or acronyms), diaphasic variants (which can be science-based variants, informal variants, and domain-based variants), dimensional variants, metonymic variants, diachronic variants, non-recommended variants, and morpho-syntactic variants. This information can be found on chapter three, where we have outlined each one in detail. This constitutes the monolingual part of the analysis.

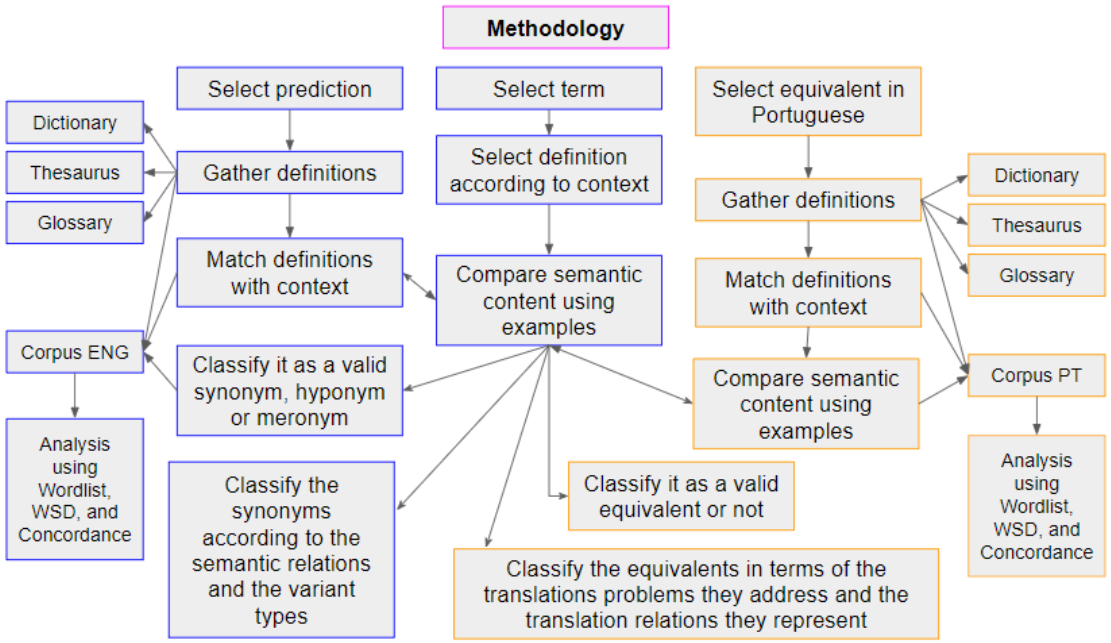
The bilingual part of our analysis involves similar steps. We intend to list the meanings of the chosen equivalents in Portuguese (see Table 7) and classify them as a valid equivalent or not considering the meanings of each term in context and its application in the retail domain. If two words share the same meaning in this common context, they will be considered suitable equivalents. Once these equivalences are judged as appropriate or inappropriate, we will classify them in terms of the

translation relations they participate in and translation problems they address, as defined by León-Araúz & Faber (2014).

The authors identified ten translation problems and they vary from lack of equivalence to partial equivalence among concepts. The translation relations, on the other hand, were defined by the authors as canonical translations, generic-specific translations, extensional translations, communicative translations, functional translations, cultural translations, descriptive translations, non-translations, and metonymic translations. In depth explanations of these translation relations and translation problems can also be found in chapter three of this work.

Figure 29 summarizes the process. The squares with a blue outline regard the monolingual part of the analysis and the squares in orange regard the bilingual stage.

Figure 29 – Methodology



Source: made by the author (2023).

As we can observe, the methodology consists of choosing the terms to be analyzed (which we already did, as Table 7 outlines), and then proceeding to the selection of the definitions and senses of these terms. The same process is applied to the predictions provided by the models and the equivalents in Portuguese. Then, these senses are allocated into the context (regarding the retail domain) and their semantic information considering this specific sense is listed. Then, we must compare this information and identify the shared aspects of senses that are evidence

of their sameness of meaning. This semantic information is the key to the classification of these relations, since it is what determines if the predictions are adequate and valid synonyms, hyponyms, meronyms and/ or equivalents for that term or not.

The intent of this chapter was to delineate how the analyses of lexical variation and lexical equivalence took place and what materials were used to accomplish these analyses. In order to do so, we described the theories and tools that supported us along the analysis stage. We have also described the terminology analyzed by us and how we gathered the data. In the next chapter, we will discuss our analysis and how we conducted the development of the methodological steps described in this chapter.

5 ANALYSES

The aim of this chapter is to lay out, depict, and describe the analyses conducted by us while using the theoretical background presented in chapters two and three to endorse and support our results. As mentioned in chapter four, our analysis is divided in two parts. One concerns the semantic and terminological analysis regarding the monolingual terminology. We will begin by evaluating the semantic content that each term carries and evokes when allocated in the context dictated by the retail domain. Once the suggested substitutes are classified as a satisfactory fit or not depending on the semantic relations attributed to them, we will move on to classifying these alignment suggestions according to the variant types proposed by León-Araúz & Faber (2014) in the hopes that this labeling helps us identify what variant type the models should be ready to predict.

The bilingual part of the analysis follows the monolingual stage and is concerned with the lexical equivalents listed for this set of terms in Portuguese. Our aim is not to classify these equivalents according to types of equivalents for this would require a theoretical commitment we did not assume on chapter three[1]. Instead, we intend to list what equivalents for the terms chosen by us exist in Brazilian Portuguese and if they fit in the retail domain or not. Moreover, we intend to evaluate which of the translation problems they would address, inspired by the translation problems elected by León-Araúz & Faber (2014).

In order to accomplish the goals of this chapter, we divided it into two sections, each regarding one of the stages mentioned in the previous paragraphs. The first one, section 5.1, addresses the monolingual analysis. The second one, section 5.2, addresses the bilingual analysis. Our aim in this chapter is to understand and exemplify the linguistic behavior and characteristics of the terminology being analyzed. Thus, the section below begins by outlying the first stage.

[1] As mentioned in chapter three, there are many different nomenclatures and definitions for lexical equivalence and each author defines equivalence according to their approach in the field of Translation Studies. Despite not defining our own concept of equivalence, we closely align with cognitive theories of meaning and will describe how this view impacts our bilingual analysis in this chapter and in chapter six.

5.1 Monolingual Analysis

This section is dedicated to the first stage of the analysis developed by us. We will begin by presenting the results found by our models and the chosen adequate choices for the terms studied. Once we have outlined the results and classified the semantic relations that appear, we will classify each occurrence according to the variant types outlined by León-Araúz & Faber (2014), which were detailed in chapter three. We will begin by tracing which of the suggestions presented by each model were accepted as potential terms for the alignment task and lexical substitution tasks considering the retail domain.

In chapter four, we described each model and listed a few terms we believed would be predicted by the models considering our previous knowledge about the terminology. As explained before, the first model, called “Roberta” considered only synonyms, while the second one, “Roberta embedding” took word embeddings into consideration as well. Finally, the last model, called “Roberta semantic frames target embedding”, considered the semantic information provided by the semantic frames while looking for predictions. It is worth to mention that this last model is the most recent one and the one that presents the best results so far. We intend to provide a few insights further on about the reason behind this.

Tables 8, 9, and 10 detail the predictions shown by each model, list the terms we tried to predict (see Table 7), and elicit the terms we regard as a good fit for the alignment and lexical substitution tasks considering the retail domain.

As we can observe, the results show a certain degree of variation and the reason behind this is the fact that each model had different contextual and linguistic features operating while the predictions were being generated. Our choice, as previously mentioned, was based on a desire to observe how the gradual implementation of linguistic information would impact the results of the models. It is important to mention that the terms considered adequate choices were classified by us taking into account our knowledge of the domain and the terminology, and the information in the corpora.

Table 8 – Roberta model

Term	Predictions	Our suppositions	Adequate choices
Plant	facility, factory, site, project, production, mill, process, complex	factory, unit, works, foundry, mill, shop, yard, industrial unit, business unit, building	facility, factory, site, mil
Site	project, facility, plant, center, campus, complex, factory	factory, unit, works, foundry, mill, shop, yard, industrial unit, business unit, building	facility, plant, center, factory
Material	manufacturing, metals, logistics, plastics, aerospace, communications, fuels, construction	stuff, product, thing, item, object, unit, commodity, artifact	metals, plastics, fuels
Article	stock, product, IPO, offering, book, issue, story, ad, business	stuff, product, thing, item, object, unit, commodity, artifact	stock, product

Source: made by the author (2023).

Table 9 - Roberta embedding model

Term	Predictions	Our suppositions	Adequate choices
Plant	factory, facility, mill, project, site, production, station, process, planting	factory, unit, works, foundry, mill, shop, yard, industrial unit, business unit, building	factory, facility, mill, site, station
Site	project, facility, website, location, campus, plant, platform, center, station	factory, unit, works, foundry, mill, shop, yard, industrial unit, business unit, building	facility, plant, center, station
Material	metal, plastic, manufacture, logistic, communication, resource, component, fuel, fabric	stuff, product, thing, item, object, unit, commodity, artifact	metal, plastic, component, fuel, fabric
Article	story, item, book, product, piece, report, essay, publication, paper	stuff, product, thing, item, object, unit, commodity, artifact	item, product, piece, paper

Source: made by the author (2023).

Table 10 - Roberta semantic frames target embedding model

Term	Predictions	Our suppositions	Adequate choices
Plant	facility, factory, site, project, mill, process, production, manufacturer, station	factory, unit, works, foundry, mill, shop, yard, industrial unit, business unit, building	facility, factory, site, mill, station
Site	location, facility, project, website, locale	factory, unit, works, foundry, mill, shop, yard, industrial unit, business unit, building	facility, locale
Material	metal, manufacture, logistic, plastic, resource, fabric, communication, component, product	stuff, product, thing, item, object, unit, commodity, artifact	metal, plastic, fabric, component, product
Article	product, item, story, piece, book, publication, report, paper, issue	stuff, product, thing, item, object, unit, commodity, artifact	product, item, piece, paper

Source: made by the author (2023).

The first step of our analysis was to define the context of each core term to be aligned and elaborate definitions for these terms. The definitions were identified based on the context we are working with, which is retail, and they were established by us considering the information provided by the company regarding each term. Therefore, these definitions were elaborated by us, taking into consideration a study of the domain and they were validated regarding the information provided by the company. These definitions come from a trustful source and are adopted by us throughout the analysis to guide our deliberations.

The definitions are explained in Table 11 along with the semantic information we identified for each term. The semantic information is the basic information necessary to comprehend the concept evoked by each term.

Table 11 – Definitions and semantic information of core terms to be aligned

Term	**Definition according to the Retail Domain**	**Core semantic Information**
Plant	A geographical location where materials are produced, or goods and services are provided. It is an industrial site (type of site) or factory where workers and machines produce goods. It can also be the nodes in a hierarchy containing further plants or it can be the grouping of several plants. It is an organization unit for dividing an enterprise according to production, procurement, maintenance, and materials planning at which quantities of products are managed.	• A geographic location • A type of site or factory • Can refer to one or more plants • A place where a certain type of activity happens
Site	A geographic location that can be a building, a group of buildings, multiple site areas, or a point within a site where either articles are produced or goods and services are provided. An enterprise resides in a site, and it can be any type of enterprise in which production activities can take place. In the system, a site can be an entire plant where you manufacture products. It is a separate, smaller facility at a plant where you manufacture a product, a specialized portion of a plant or facility, etc. It is used to communicate place data in the business process. It can also be an indicator of the town or district in which the business entity is situated.	• A geographic location • Can have one or more buildings • A place where enterprises reside • A type of plant • Part of a plant
Material	An object that includes products, materials,	• An object or

	articles, and services and that is the subject of business activity. The material can be traded, used in manufacture, consumed, or produced. It is a substance or object dealt with on a commercial basis or used, consumed, or generated during production. It can be a chemical element or a compound. It can be a specific part of the product to be assembled or the entire product. It can be a group of products according to their attributes or a single unit. It can be a retail sector.	service subject to business activity • Hypernym term that englobes other objects • Component used to make an article
Article	An object subject to commercial transactions that includes products, materials, articles, and services. It is usually ordered for a site and sold. It can also be the smallest unit or customer pack.	• An object or service subject to business activity • Part of an object

Source: made by the author (2023).

These definitions and the semantic information encoded by each one of the terms was what we took into account when defining what predictions would be a satisfactory fit in the term alignment and lexical substitution stages and in the translation stage. The criteria used to determine what constitutes an adequate term choice take into consideration the context and the semantic information we have outlined (see Table 11). Therefore, in order to be a valid option, the term suggested by the model must share the semantic information that the target term has, it must have a similar sense, and this sense must fit the context without altering the information being represented by the target term. It is important to mention that our analysis was extremely strict and terms which shared some but not the totality of semantic information were not considered synonymous. We have also outlined terms that have other semantic relations in common, such as hyponymy or meronymy.

Our definitions of these semantic relations align with the postulates made by Apresjan (1974), Murphy (2003, 2010) and Croft & Cruse (2004) which were discussed in chapters two and three of this work. In short, we regarded polysemy in

our analysis as a phenomenon that happens when the target term and the suggested prediction share the same meaning and the same semantic information outlined by us in Table 11. Hyponymy and hypernymy when considered when one term is a type of another. And meronymy was taken into consideration when a term is part of another. The examples below summarize our vision of these semantic relations:

- Synonymy: couch – sofa (sameness of meaning)
- Hyponymy: cat – animal (type of relation)
- Meronymy: finger – hand (part of relation)

Although semantic relations are central to the scope of this work, they are not the only linguistic features considered in our theoretical background. We would like to recall the variant types defined by León-Araúz & Faber (2014), since they will be used later in the analysis, for we intend to classify the suggested variants according to this categorization. As defined by the authors, there are nine types of variants:

- Orthographic variants: as the name states, variants in which the term is the same but spelled differently due to reasons that can be geographical or not.
- Diatopic variants: variants that change according to culture- specific terms or dialects.
- Short form variants: can be abbreviations or acronyms.
- Diaphasic variants: can be science-based, informal, or domain-based variants.
- Dimensional variants: usually multi-word terms that convey different dimensions of the same concept.
- Metonymic variants: based in the semantic relation, terms are organized according to their parts.
- Diachronic variants: variants affected by the passage of time.
- Non-recommended variants: offensive or inadequate terms.
- Morpho-syntactic variants: as the name states, variants in which the terms have a different component or a different order.

Considering this line of evaluation, we will begin by analyzing these similarities of meaning and senses to identify lexical relations. Then, we will classify the predictions offered by each model according to the categorizations identified by León-Araúz & Faber (2014). In order to do so, this section is divided according to each model to be analyzed. Therefore, subsection 5.1.1 will be concerned with the

results provided by the Roberta model, subsection 5.1.2 regards the results provided by Roberta embedding model, and subsection 5.1.3 allows us to consider the results of Roberta semantic frames target embedding model.

5.1.1 Roberta

This first model took into consideration the context provided for each term and gave us predictions based on the synonyms available in the dataset. In terms of linguistic information, this is a simple model with very few features to draw information from. Overall, the results presented by this model were not accurate enough for its performance to be considered satisfactory and most of the predictions of the model were not valid options for lexical substitution. Figure 30 outlines the results provided by this model.

Figure 30 – Results of model Roberta

Target	Context	Predictions
Plant	The 140,000-square-foot **plant** is a testament to advanced manufacturing technology.	facility, factory, site, project, production, mill, process, complex
Site	The inauguration of the **site** in Cajamar, the biggest and most advanced integrated cosmetics research, production and distribution center in South America, was a huge success.	project, facility, plant, center, campus, complex, factory
Materials	The solutions for innovation in **materials**, commercial and logistics models, as well as cross-cutting proposals involving more than one topic, were analyzed.	manufacturing, metals, logistics, plastics, aerospace, communications, fuels, construction
Article	The company's newest **article** is selling well, according to the reports we received.	stock, product, IPO, offering, book, issue, story, ad, business

Source: made by the author (2023).

Considering the individual pairs of terms to be aligned, some had better results than others. “Plant” in particular, represents a part of the model that worked pretty well if compared to the other terms, with 50% of valid options considering the

context. The statistics regarding valid lexical substitution options for “site”, “material”, and “article” were under the 50% margin of results.

Firstly, we believe that the suggestions for “plant” had a superior performance due to the context. The model did not suggest terms related to the botanic sense of the term “plant”, which is a positive outcome which shows that the model was able to disambiguate the term “plant”. On the contrary, the terms that it suggested as a replacement fit the context most of the time or had some connection to the meaning of the target term in the Retail Industry. The Roberta predictions considered a good fit for “plant” are “facility”, “factory”, “site”, and “mill”. As we can observe, the correct substitution term was among these predictions, which was another positive surprise since we were not expecting a domain-specific term such as “site” to appear. Once again, we believe the context had an important role to play when it comes to this result.

“Facility” was considered a good fit for the substitution task due to the following definitions:

- a place, amenity, or piece of equipment provided for a particular purpose. Example: "cooking facilities".
- something designed, built, installed, etc., to serve a specific function affording a convenience or service. Example: “educational facilities, new research facility”.
- buildings, pieces of equipment, or services that are provided for a particular purpose. Example: “What recreational facilities are now available?”.
- a facility is something such as an additional service provided by an organization or an extra feature on a machine which is useful but not essential. Example: “It is very useful to have an overdraft facility”.

As we can observe, these definitions of “facility” match the semantic information provided by “plant” in the sense that both refer to a geographical location where certain types of business and/or production related activity happens. Despite having other senses which do not fit the criteria, such as “something that permits the easier performance of an action” and “an ability to do or learn something well and easily”, facility seems to be a good candidate for term alignment in this case, especially because it does not seem to corrupt the sense of “plant” in the retail domain. Thus, when taking into account the definition of synonymy posed by the authors used in

this work, “facility” was considered a synonym of “plant” by us. “Factory” is another term with many shared senses, as exemplified below:

- a building or group of buildings where goods are manufactured or assembled chiefly by machine. Example: "a clothing factory".
- a person, group, or institution that continually produces a great quantity of something specified. Example: "a huge factory of lying, slander, and bad English".
- a building or set of buildings where large amounts of goods are made using machines. Example: “a car factory”.
- a station where factors reside and trade. Example: “a colonial factory”.
- a building or set of buildings with facilities for manufacturing. Example: “a production building”.

These definitions of “factory” clearly denote the existence of a building or facility in which activities related to the retail industry take place, especially when these activities are related to the manufacturing and production process. Usually, the manufactured products are intended to be the goods of a commercial transaction. If we refer back to the definition of “plant” posed by us (and endorsed by our study of the Retail domain and the material provided by the company), we can observe that one can refer to “plant” as a “[…] factory where workers and machines produce goods” (see Table 11). “Factory” seems to have much more shared senses with “plant” than “facility” does. These senses, however, seem to be much more specific and detailed. While “plant” can be used to refer to a geographical location in which any activity related to retail can take place (even stocking goods), “factory” refers to places in which the only activity that takes place is the production of these goods. Thus, we classified “facility” as a hyponym of “plant”, due to its “type_of” semantic relation.

“Site” is the next term we must analyze. We know it is the correct option for term substitution in the retail domain. Thus, the fact that it was suggested by the model is an extremely positive outcome. These are the definitions we gathered for “site”:

- an area of ground on which a town, building, or monument is constructed. Example: "the proposed site of a hydroelectric dam".
- a place where a particular event or activity is occurring or has occurred. Example: "the site of the Battle of Antietam".

When compared to the definition of “site” used in this study (see Table 11), these definitions are poor in terms of semantic content. They refer to a geographical location where an activity occurs, however, this activity is not specified nor detailed. Yet, even with few information regarding the term, it is classified as a synonym for “plant” due to the fact that it is its counterpart when it comes to alignment and substitution tasks in NLP.

The last term to be analyzed is “mill”. These are the two definitions that were closest to the context we are working with:

- a place that processes things or people in a mechanical way. Example: "a correspondence school that was just a diploma mill".
- a building fitted with machinery for a manufacturing process. Example: "a steel mill".

Although “mill” usually refers to a place equipped with machinery for grinding grain into flour, we can observe that one of its definitions can be considered a metaphorical way to refer to things made mechanically, without a personalized or unique feature. The other sense could refer to a building with machinery that does not necessarily concerns the process of turning grain into flour. Thus, in this particular case, it could be a candidate for term alignment due to the fact that there are shared semantic information between “mill” and “plant”. Similar to “factory”, “mill” seems to have a semantic relation of “type_of” with “plant”, since it is specifically concerned with the manufacturing process. Thus, it was classified as a hyponym instead of a synonym. When it comes to “plant”, the predictions of the first model were adequate and the correct term – “site” – was among the answers.

We now move on to the analysis of “site”. The terms that could be considered adequate fits in the alignment task were “facility”, “plant”, “center”, and “factory”. In this case, once again, the model predicted the correct answer, which is “plant”. Both “facility” and “factory” were analyzed above for they were predictions for “plant” as well. Their definitions do not change due to the sameness of context. Therefore, “facility” remains as a synonym and “factory” is still classified as a hyponym. When it comes to “plant”, the definition found is displayed below:

- a place where an industrial or manufacturing process takes place. Example: "the company has 30 plants in Mexico".

As we can observe, this definition is not very specific considering the nature of the term in context. The definition of “plant” used in this work (see Table 11) is much more complete than the one we could find. This reveals a lack of specialized knowledge in the sources we used to gather the definitions. When compared to the definitions elaborated by us, this definition lacks a few core semantic information such as the characteristics of the place (it can be one or more buildings, parts of a building) and the characteristics of the activities that take place at the location (which ranges from a service being provided to the production and storage of goods). Despite a much less detailed explanation of the “plant” term, it was considered a synonym for “site” due to the fact that it is indeed the correct option in this lexical substitution task.

When it comes to “center”, there was also only one definition which would match the context. It is exemplified below:

- a place or group of buildings where a specified activity is concentrated. Example: "a center for medical research".

In this case, the information regarding the location seems very much like the definition of “site” and the activity is not specified, which could lead one to use the term to refer to a site, since there are many activities that could be executed in a center. Therefore, “center” was also classified as a synonym by us. Thus, among the predictions of “site”, the Roberta model provided us with three synonyms (one of them being the correct term) and one hyponym.

Next, we have the term “material” to be analyzed. The considered predictions for this term were “metal”, “plastic”, and “fuel”. As we can see, all of these terms refer to types of materials that can be used to make other products. However, they could still be subject to a commercial transaction. Thus, we will consider their definitions in order to classify them as hyponyms and not synonyms. “Metal”, the first of our selected predictions, has the following definition that could match the context:

- a solid material that is typically hard, shiny, malleable, fusible, and ductile, with good electrical and thermal conductivity (e.g., iron, gold, silver, copper, and aluminum, and alloys such as brass and steel). Example: "being a metal, aluminum readily conducts heat".

As we can observe, this definition classifies a metal as a type of material and proceeds to give characteristics of the substance. This definition matches one of the

semantic features numerated by us for “material”, which is the ability of being a component used to make an article. Additionally, one of the characteristics of “material” according to our definition is that it could be used as an umbrella term that englobes other objects. This seems to be the case considering “metal”. Consequently, “metal” was classified as a hyponym of “material” in our analysis. “Plastic” has a similar definition and according to the corpus it could be described as:

- a synthetic material made from a wide range of organic polymers such as polyethylene, PVC, nylon, etc., that can be molded into shape while soft and then set into a rigid or slightly elastic form. Example: "bottles can be made from a variety of plastics".

Similar to “metal”, this term was classified as a hyponym of “material” due to the fact that it is a type of material, as stated by its definition in this context. The same thing happened with “fuel”. It can be described as a “material such as coal, gas, or oil that is burned to produce heat or power”. Once again, the definition itself classifies “fuel” as a type of material. Thus, the three terms predicted by this model are hyponyms due to the fact that they could be purchased or sold. It is important to notice that the Roberta model did not suggest the term “article” as a synonym for “material”, which would be the correct answer.

Lastly, we must analyze the predictions for “article”. There were two selected terms in this case: stock and product. The definitions for “stock” that would match the context are listed below:

- the goods or merchandise kept on the premises of a business or warehouse and available for sale or distribution. Example: "the store has a very low turnover of stock".
- the raw material from which a specified commodity can be manufactured. Example: "the fat can be used as soap stock".

The first listed definition of “stock” is very similar to our definite definition of “article” (see Table 11) in which we classified it as an object or service subject to business activity or even part of an object. Thus, if we consider this first definition of “stock”, the term could be considered a synonym of “article”, since it refers to a significant number of articles. The second definition goes to show us some similarities with the “material” semantic features. However, the model did not consider it an appropriate suggestion for “material” and instead considered it only for

“article”, which is not entirely wrong. The definitions of the other valid suggestion “product” can be found below:

- an article or substance that is manufactured or refined for sale. Example: "food products".
- commercially manufactured articles, especially recordings, viewed collectively. Example: "too much product of too little quality".

In this case, both definitions share important semantic information with the term “article”, taking into consideration the definition of “article” in the retail industry. Therefore, “product” can also be considered a synonym of “article” taking the context in consideration. Despite the fact that Roberta model predicted two valid synonyms for “article”, none of them was the correct one for the substitution task, which in this case would be “material”. Table 12 summarizes the results of this model.

In regard to this part of the analysis, we can affirm that the first model performed well when it comes to the first pair of terms since it suggested that “plant” and “site” are synonyms. Additionally, the model understood that the context excluded the botanic sense of “plant” and the computational sense of “site”, since predictions related to these senses did not appear. The same was not the case for the second pair.

“Material” can be considered an umbrella term because it is a hypernym used to address a multitude of objects, components, and elements that compose a certain thing. Thus, the model suggested a variety of replacement terms to be aligned, yet just a few were chosen by us and none of them were classified as synonyms. Most of these were not a good fit for this particular context, although they could work very well if “material” were referring to more specific elements. Finally, “article” was mainly taken by the model as a piece of writing. Therefore, most of the suggestions were not a good fit for the context being used. Despite understanding the contexts related to “plant” and “site”, the model seems to lack an understanding of the terms “article” and “material”.

Table 12 – Roberta results

Target term	**Selected predictions**	**Semantic relations**
Plant	Facility	Synonymy
	Site	
	Factory	Hyponymy
	Mill	
Site	Facility	Synonymy
	Plant	
	Center	
	Factory	Hyponymy
Material	Metal	Hyponymy
	Plastic	
	Fuel	
Article	Stock	Synonymy
	Product	

Source: made by the author (2023).

These remarks conclude this part of our analysis. We must now classify each variant according to the variant types posed by León-Araúz & Faber (2014) which are orthographic variants, diatopic variants, short form variants, diaphasic variants, dimensional variants, metonymic variants, diachronic variants, non-recommended variants, and morpho-syntactic variants. We can exclude the following variant types:

- orthographic variants: there were no orthographic variations among the predictions presented by the model.
- diatopic variants: there were no cultural or dialectical variants among the predictions.
- short form variants: there were no abbreviations or acronyms among the predictions.
- dimensional variants: there were no multi-word variants among the predictions.
- diachronic variants: there were no old or archaic terms among the predictions.
- non-recommended variants: there were no terms with a negative connotation among the predictions.

- morpho-syntactic variants: there were no variants with morpho-syntactic differences among the predictions.

Thus, we are left with diaphasic variants and metonymic variants to analyze. Although we were able to classify all the predictions according to semantic relations, we did not identify meronymy in this case. We did not identify predictions which had the "*part_of*" type of relationship with the terms chosen in our analysis. Hence, we are left with diaphasic variants which can be of three types, according to León-Araúz & Faber (2014): science-based variants, informal variants, and domain-based variants. The predictions cannot be classified as science-based variants because there were no scientific nomenclatures, jargons, formulas, or symbols among the predictions. According to the author, this type of variant would be the default term choice in a specialized scenario. However, we can observe that it is not the case in this particular domain. Additionally, none of these predictions can be considered informal variants because there were no colloquial variants or generic variants.

However, when it comes to domain-based variants, we believe that all of the predictions were based on this type of information. This type of variant affects the semantics and the communicative situation, especially if term preferences change across specialized domains. Due to the fact that these variants are related to what constitutes specialized terminology and what does not, we have classified all the predictions provided by Roberta model as domain-based variants, because they are rooted on what is part of the retail domain and what is not. This corroborates the hypothesis we raised on chapter three of this work, when we stated our beliefs that this type of variant would be the most common.

These reflections on semantic relations and the variant types classification close our analysis of the first model. Once again, we reiterate that this model was the first one we executed and lacked some important linguistic features for the predictions. Next, we are going to evaluate Roberta embedding, which can be considered an intermediate model if compared to the ones we have developed. As the name suggests, it used word embeddings on its execution.

5.1.2 Roberta Embedding

This model took into consideration the context provided for each term and gave us predictions based on the synonyms and word embeddings available in the

dataset. In general, the results presented by this model were not so different from the previous model, although there was a bigger number of valid answers than the Roberta model. Figure 31 presents us the results of the model.

Figure 31 – Results of Roberta embedding model

Target	Context	Predictions
Plant	The 140,000-square-foot **plant** is a testament to advanced manufacturing technology.	factory, facility, mill, project, site, production, station, process, planting
Site	The inauguration of the **site** in Cajamar, the biggest and most advanced integrated cosmetics research, production and distribution center in South America, was a huge success.	project, facility, website, location, campus, plant, platform, center, station
Materials	The solutions for innovation in **materials**, commercial and logistics models, as well as cross-cutting proposals involving more than one topic, were analyzed.	metal, plastic, manufacture, logistic, communication, resource, component, fuel, fabric
Article	The company's newest **article** is selling well, according to the reports we received.	story, item, book, product, piece, report, essay, publication, paper

Source: made by the author (2023).

When it comes to these predictions of each individual term, the results varied slightly. “Site” in this case, represents the term that benefited the most when it comes to the valid predictions, since the results were satisfactory. Once again, terminology related to other senses outside the scope of the context were not suggested by this model for the terms “plant” and “site”. The same did not happen for “article” and “material” which seems to be a challenging pair for the models so far.

The Roberta embedding predictions considered a good fit for “plant” are “facility”, “factory”, “site”, “mill”, and “station”. As we can observe, the correct prediction for the alignment task in this case, which is “site”, made a happy appearance in these suggestions as well. Since we have already analyzed the terms “facility”, “factory”, “site”, and “mill”, we will not detail them a second time. These terms were classified, respectively, as a synonym, a hyponym, a synonym, and a hyponym. Considering the term “station”, which has not been analyzed yet, we identified the following sense when it comes to the context at hand:

- a place or building where a specified activity or service is based. Example: "a research station in the rainforest".

As we can observe, the definition above refers to a geographical location and the activities that take place at this location are not specified. Referring back to the definition of “plant” posed by us and used broadly in the retail domain (see Table 11), we are able to classify “station” as a synonym of “plant”, due to the usability it has in this context. The term could be used to refer to a “production station”, for example. Hence, Roberta embedding model provided us three synonyms and two hyponyms for “plant”.

Next, we must analyze the valid results for “site”. The predictions considered valid by us are the following: “facility”, “plant”, “center”, and “station”. All of these terms have been analyzed by us previously and since they refer to the same concept, we will not go over the analysis twice. Thus, Roberta embedding provided us with four synonyms for “site”, one of them being the correct option for this alignment task.

When it comes to “material”, the results were the terms “metal”, “plastic”, “component”, “fuel”, and “fabric”. “Component” and “fabric” are the terms we have not analyzed thus far. “Metal”, “plastic”, and “fuel” have been classified as hyponyms and therefore will not have further clarifications dedicated to them. Regarding “component”, we have identified the following definition:

- a part or element of a larger whole, especially a part of a machine or vehicle. Example: "stereo components".

In this case, as the definition clearly states, we are handling a “*part_of*” type of semantic relation, instead of a “*type_of*” relation. Thus, according to the definitions posed by Murphy (2010) and Croft & Cruse (2004), we have classified “component” as a meronym of “material”, instead of a synonym or a hyponym. Although we see why “component” could fit a context in which the term being referred to is a part of a “material”, this is not the correct answer for this case. Lastly, we gathered the “fabric” definition, stated below:

- cloth or other material produced by weaving or knitting fibers. Example: "heavy silk fabric".

This definition, as we examine, refers to a type of material due to the fact that it would be a good lexical substitute only in very specific contexts. This type of context

includes the fashion industry or the retail domain only when it concerns the commercial transaction of clothes or items that can be wore. If the context does not concern these listed conditions, then “fabric” does not match “material” in terms of definition and shared semantic features. Therefore, “fabric” was classified as a hyponym of “material” which means that Roberta embedding model provided us four hyponyms, one meronym, and no synonyms for “material”. The correct answer “article” was not among the predictions of the model.

Lastly, we must analyze the predictions offered by the model when it comes to “article”. The valid predictions provided by Roberta embedding model in this case are “item”, “product”, and “piece”. “Product” is the only term that has already been analyzed by us and it was classified as a synonym. Consequently, will no longer be addressed. “Item”, the first valid suggestion to be analyzed by us, can be defined as

- an individual article or unit, especially one that is part of a list, collection, or set. Example: "the items on the agenda".

This definition classifies item as an article, but it does not specify the usability, or the functions performed by this item. Despite this definition not being a complete one, we understand that the term “item” can be used to refer to a variety of things (including products destined to commercial transactions) and trying to define each one of the activities an item could be submitted to would probably fall short and narrow its usability. We believe that an article could be referred to as an item destined to sales or production. Thus, we have classified “item” as a synonym for “article”, despite it not being the correct answer in this case.

Next, we turn our attention to “piece”, which has the following definitions:

- a portion of an object or of material, produced by cutting, tearing, or breaking the whole. Example: "a piece of cheese".
- an item of a particular type, especially one forming one of a set. Example: "a piece of luggage".

If we take into consideration our definition of “article”, we can observe that it is described as “an object subject to commercial transactions that includes products, materials, articles, and services. […] it can also be the smallest unit or costumer pack”. Thus, “piece” classifies as a synonym in this case, even if it looks very much like a meronym. Hence, the Roberta embedding model provided us with three

synonyms for “article”, however, none of them were the correct answer, which would be “material”.

These remarks conclude our analysis of the prediction provided by Roberta embedding model. We elaborated Table 13 which summarizes the results of this model.

Table 13 – Roberta embedding results

Target term	**Selected predictions**	**Semantic relations**
Plant	Facility	Synonymy
	Site	
	Station	
	Factory	Hyponymy
	Mill	
Site	Facility	Synonymy
	Plant	
	Center	
	Station	
Material	Metal	Hyponymy
	Plastic	
	Fuel	
	Fabric	
	Component	Meronymy
Article	Item	Synonymy
	Product	
	Piece	

Source: made by the author (2023).

In regard to this part of the analysis, we can affirm that this model also performed well when it comes to the first pair of terms since it suggested that “plant” and “site” are synonyms. Additionally, this model was also able to understand that the context excluded the botanic sense of “plant” and the computational sense of “site”, since predictions related to these senses were a minority – there was one suggestion for “plant” taking the botanic sense into consideration (planting) and one suggestion for “site” taking the computational sense into account (website).

“Material” and “article”, on the other hand, are proving to be a challenge for the models analyzed so far. “Material” was once again considered an umbrella term due to its hypernymic nature. The model suggested a variety of replacement terms to be substituted, yet just a few were chosen by us and none of them were classified as synonyms. Once again, most of these suggestions were not a good fit for this particular context, although they could work if “material” were referring to more specific elements. Despite the fact that a new semantic relation appeared (meronymy), it is important to remember that the correct answer was not among the predictions made by the model.

Finally, “article” was once again mainly taken by the model as a representation of a piece of writing. Therefore, most of the suggestions were not a good fit for the context being used. Hence, we conclude that this model, despite using extra linguistic features to look for lexical substitutes for these terms, did not necessarily provide better results, although it provided a bigger number of valid predictions that we considered in our analysis.

We now move on to the classification of each variant according to the variant types posed by León-Araúz & Faber (2014). Due to the similarities between model results, we can exclude the same variant types for the same reasons, as the following list exemplifies:

- orthographic variants: there were no orthographic variations among the predictions presented by the model.
- diatopic variants: there were no cultural or dialectical variants among the predictions.
- short form variants: there were no abbreviations or acronyms among the predictions.
- dimensional variants: there were no multi-word variants among the predictions.
- diachronic variants: there were no old or archaic terms among the predictions.
- non-recommended variants: there were no terms with a negative connotation among the predictions.
- morpho-syntactic variants: there were no variants with morpho-syntactic differences among the predictions.

Thus, we are once again left with diaphasic variants and metonymic variants to analyze. Contrary to what happened with the results provided by Roberta model, this model provided us with one prediction which was classified as a meronymy. It is the case of “component”, which was classified as a meronym of “material”. According to León-Araúz & Faber (2014), the metonymic variants are based in this type of semantic relation, the one that designates the concept according to its parts. Thus, we conclude that “component” is a metonymic variant.

Diaphasic variants, the remaining variant type, is divided into science-based variants, informal variants, and domain-based variants. The Roberta embedding predictions cannot be classified as science-based variants because there were no scientific nomenclatures, jargons, formulas, or symbols among the predictions, as in the previous model. In a similar way to the Roberta predictions, none of the Roberta embedding model suggestions can be considered informal variants because there were no colloquial variants or generic variants.

Thus, with the exception of the term “component”, all of the other predictions posed by Roberta embedding fall into the domain-based variants type of category. Because we can observe the similarities between the results provided by these two models, we reiterate that these variants are related to what constitutes specialized terminology and what does not. Thus, as explained by León-Araúz & Faber (2014), they affect the semantics and the communicative situation and must be taken into account when dealing with the specialized terminology field.

These reflections close our analysis of the Roberta embedding model. This model is what we consider an intermediate model in terms of linguistic information that it takes into consideration when in execution. Next, we are going to evaluate our most recent model, called Roberta semantic frames target embedding. The next section is dedicated to the analysis of its results.

5.1.3 Roberta semantic frames target embedding

As the name suggests, this model takes into consideration the following linguistic information: synonymy, word embedding, and semantic frames. Unfortunately, the addition of semantic frames as an input for the model to work with did not seem to improve the results significantly when it comes to these four terms being analyzed. Despite having noticed in other tests that the addition of frame semantics improves

the results provided by the model, this was not observable in this case. Figure 32 outlines the results of this model and the semantic frames used by it.

Figure 32 – Results of the Roberta semantic frames target embedding model

Target	Context	Semantic Frame	Predictions
Plant	The 140,000-square-foot **plant** is a testament to advanced manufacturing technology.	*locale_by_event_scenario*	facility, factory, site, project, mill, process, production, manufacturer, station
Site	The inauguration of the **site** in Cajamar, the biggest and most advanced integrated cosmetics research, production and distribution center in South America, was a huge success.	*locale_by_event_scenario*	location, facility, project, website, locale
Materials	The solutions for innovation in **materials**, commercial and logistics models, as well as cross-cutting proposals involving more than one topic, were analyzed.	*manufacturing_scenario*	metal, manufacture, logistic, plastic, resource, fabric, communication, component, product
Article	The company's newest **article** is selling well, according to the reports we received.	*manufacturing_scenario*	product, item, story, piece, book, publication, report, paper, issue

Source: made by the author (2023).

Among the predictions for “plant” we can still find the correct answer “site”. However, “plant” was not predicted as a substitution option for “site” by this model. The other predictions reflect the same results observed by us in the previous two models. The addition of the semantic frame did not seem to alter the results significantly in this case and the pair of terms “material” and “article” continues to pose a challenge.

When it comes to the first term to be analyzed, which is “plant”, this model predicted numerous replacement items and we identified the following as valid options: “factory”, “facility”, “site”, “mill”, and “station”. All of them have already been analyzed due to the fact that they have been predicted by the previous models as well. They were classified respectively as a hyponym, a synonym, a synonym, a hyponym, and a synonym.

“Site”, on the other hand, had two selected valid answers and none of them was the correct term. These predictions are “facility” and “locale”. “Facility” is another analyzed term and it was previously classified as a synonym. When it comes to “locale”, we have identified the following definition:

- a place where something happens or is set, or that has particular events associated with it. Example: "her summers were spent in a variety of exotic locales".

As we can observe, this definition takes into account a description of the activities that happen at this location and the place where something is set, without accounting for what type of place it is or what type of activity takes place at this location. It seems like this lexical unit requires a type of event linked to it, which could be interpreted as some sort of retail activity. Thus, considering this definition of “locale” and the definition of “site” posed by us (see Table 11), we have classified the term “locale” as a synonym of “site”.

Regarding “material”, we have identified five valid terms for the substitution task. They are “metal”, “plastic”, “fabric”, “component”, and “product”. “Metal”, “plastic”, and “fabric” were classified as hyponyms of the term “material”. Meanwhile, “component” is classified as a meronym and “product” is classified as a synonym. It is worth noting that the correct answer was not among the predictions.

Lastly, the considered valid options for “article” were “product”, “item”, and “piece”. All of them were classified as synonymous lexical units for “article”. The addition of semantic frames did not help the Roberta semantic frames target embedding model understand that the referred sense of “article” was not a textual one. Thus, the quality of the predictions in this case was similar to the other models. Table 14 summarizes these results.

Despite these results, it is important to state that when it comes to this last model, which is the most sophisticated one when it comes to the quality of linguistic information, the model performed well when being executed with other terms from the selected dataset. Although we recognize that four terms are not nearly enough to measure the efficiency of a model such as this one, all of our statements regarding the performance of this model were made taking into account solely this analysis. The comments regarding the previous models also considered exclusively the analyses made here.

Table 14 – Roberta semantic frames target embedding results

Target term	**Selected predictions**	**Semantic relations**
Plant	Facility	Synonymy
	Site	
	Station	
	Factory	Hyponymy
	Mill	
Site	Facility	Synonymy
	Locale	
Material	Metal	Hyponymy
	Plastic	
	Fabric	
	Component	Meronymy
	Product	Synonymy
Article	Item	Synonymy
	Product	
	Piece	

Source: made by the author (2023).

In order to classify the selected terms according to the variants proposed by León-Araúz & Faber (2014), we once again ruled out orthographic variants, diatopic variants, short form variants, dimensional variants, diachronic variants, non-recommended variants, and morpho-syntactic variants. In this case, we once again categorized “component” as a metonymic variant, due to the fact that it was classified as a meronym of “material”. The other variants were classified as diaphasic variants, more specifically the domain-based type of diaphasic variants. Even if some of the predicted terms are not necessarily from the scope of specialized domain, they do not fit the criteria to be classified as any other type of variant. Additionally, the terms were classified as valid options or invalid options according to the preferences of the domain we have been working with. Thus, it makes sense to classify these terms according to this type of variant.

These classifications according to the semantic relations among terms and considering the variant types defined by León-Araúz & Faber (2014) close our remarks of the monolingual analysis. In terms of the predictions, the results were

satisfactory in certain aspects of evaluation, but lacked precision in others. The overall results were good and special attention must be given to the pair “plant” and “site”, which undoubtedly posed a challenge that the models were able to overcome. The fact that the majority of the predictions was not related to the botanic sense of “plant” and the computational sense of “site” is considered an extremely positive point by us.

It is important to note that the models still do not point at one certain prediction for each term. Instead, they offer us a variety of terms and we sort it out by classifying them as a valid option or an invalid option for substitution. Thus, these models still require a preliminary human interference in the sorting of these predictions in order to give a satisfactory set of answers. First, we ruled out all the terms that did not have shared senses or semantic information related to the term in context. This was our first interference. Second, we classified each of the remaining terms according to its semantic relation to the target term. Then, we classified these predictions according to the variant types identified by León-Araúz & Faber (2014). Finally, with our knowledge of the domain, we were able to pin out the correct alignment pairs.

Our aim when developing these models was to create a computational tool that is able to point to a specific term in the domain and context provided and have this prediction be the correct one. Perhaps trying to develop a completely automatized model for lexical substitution is a little too bold and ambitious, but it is a goal we could keep in mind for the long-distance future. Undoubtedly, there are still improvements to be made. The objective of including semantic frames into the model was to help with word sense disambiguation and help enhance the quality and precision of the predictions.

The last model analyzed (Roberta semantic frames target embedding model) is the most recent model we have developed, and it includes the semantic frames. Despite the results not being the best ones in the analysis of this work, we could observe improvements in the totality of terms used as target words for this model. One of the reasons behind this result considering the terms “plant”, “site”, “material”, and “article” could be the source of the semantic frames. The platform used as a source of semantic frames was FrameNet – which was detailed in chapter two of this work – and it is not a tool concerned with specialized terminology. Additionally, the frame choices made by us could be equivocated, which would have impacted

negatively in our results, especially because we know the importance of specialized, domain specific knowledge when it comes to the development of successful models.

Another reason for this could be that perhaps synonymy, word embeddings, and semantic frames are simply not enough linguistic information for the models to draw information from. More about these suppositions will be addressed in the final considerations chapter. We now close this first part of our analysis and move on to the bilingual stage of investigation.

5.2 Bilingual Analysis

This section of our analysis regards the part of the study in which we address lexical equivalence. Our aim is to identify the equivalents of the four terms used in this analysis in Brazilian Portuguese. We must take into account the context, the translation problems we may face when handling specialized terminology and translation, and the resources we have to perform these translations. Since our model is a lexical substitution kind of model and not a machine translation-based model, we will no longer refer to the predictions posed by them. Instead, we will divide this section into two subsections to analyze each pair of terms individually and their equivalences.

Since we did not have a model to provide us with equivalent predictions for Brazilian Portuguese for each term, we came up with a list of possible lexical equivalents for these terms and displayed them on Table 7, which can be found in chapter four. These predictions were elaborated as follows:

- Material & article: *material, artigo, produto, coisa, item, objeto, unidade, mercadoria, artefato.*
- Plant & site: *planta, site, fábrica, unidade, fundição, loja, unidade industrial, unidade de negócio, prédio.*

Then, we used these predictions to look for definitions for each term in Portuguese. We also used the compiled corpus as a source of information and specialized dictionaries and thesaurus we found online. This led us to other equivalence options, and we included them in our equivalent alternatives. The results of this first analysis can be found in Figure 33.

Figure 33 – Bilingual results

Target	Context	Equivalent predictions	Valid equivalent
Plant	The 140,000-square-foot **plant** is a testament to advanced manufacturing technology.	Planta, site, fábrica, unidade, fundição, loja, unidade industrial, unidade de negócio, prédio	fábrica, departamento, loja, loja física, loja âncora, varejo, unidade industrial, unidade de negócio, canal de distribuição, fonte de suprimentos, estoque
Site	The inauguration of the **site** in Cajamar, the biggest and most advanced integrated cosmetics research, production and distribution center in South America, was a huge success.	Planta, site, fábrica, unidade, fundição, loja, unidade industrial, unidade de negócio, prédio	fábrica, departamento, loja, loja física, loja âncora, varejo, unidade industrial, unidade de negócio, canal de distribuição, fonte de suprimentos, estoque
Materials	The solutions for innovation in **materials**, commercial and logistics models, as well as cross-cutting proposals involving more than one topic, were analyzed.	Material, artigo, produto, coisa, item, objeto, unidade, mercadoria, artefato	artigo, produto, objeto, unidade, mercadoria, bem intermediário, bem de comparação, bens de conveniência, commodity, estoque, granel
Article	The company's newest **article** is selling well, according to the reports we received.	Material, artigo, produto, coisa, item, objeto, unidade, mercadoria, artefato	artigo, produto, objeto, unidade, mercadoria, bem intermediário, bem de comparação, bens de conveniência, commodity, estoque, granel

Source: made by the author (2023).

As we can observe, there are multiple terms in Brazilian Portuguese to analyze. We will divide this section into two subsections to fully examine each term and we will provide translations for the definitions in Portuguese we choose to use in this part of the analysis. Our classification of each term as a valid lexical equivalent relies on the view of equivalence described on chapter three. According to what we have defined in our theoretical background, we assumed a theorical compromise with cognitivism. It means that the equivalents and the terms in English must work as an access point to the same conceptual structure and bear the same semantic information.

We are going to classify the lexical equivalents at the end of each analysis according to the translation problems and the translation strategies defined by León-Araúz & Faber (2014). As we could observe in our theoretical background, the authors define ten translation problems when it comes to cross-lingual senses. These translations problems were described as follows:

- Translation problem number 1: the entity exists in both cultures, but the term for it in one language culture is more general or more specific than the other.
- Translation problem number 2: the entity exists in both cultures, but only one language culture has a term for it.

- Translation problem number 3: the entity exists in both cultures, yet the terms are not exact correspondents because they highlight different aspects of the concept or focus on it from different perspectives.
- Translation problem number 4: the entity exists in both cultures and both language cultures have terms for it, but only in one language the concept has been lexicalized in several variants with different communicative or conceptual differences.
- Translation problem number 5: the entity exists in both cultures and both language cultures have terms for it, which approximately correspond. However, the lexical categories appear to have different structures in each culture and thus seem to operate on different design principles.
- Translation problem number 6: the entity exists in both cultures, but its cultural role in each one is different. This leads to a conceptual mismatch and lack of correspondence.
- Translation problem number 7: the entity exists in only one of the cultures, but its name has been adopted in the other culture to refer only to the foreign culture-specific concept.
- Translation problem number 8: the entity exists in both cultures, but one culture has recycled a term from the other culture to refer to another totally different concept.
- Translation problem number 9: the entity exists in only one culture and is totally unknown in the other without any designation.
- Translation problem number 10: the entity exists in both cultures, but one of the cultures may refer to it with a metonym designation and be ambiguous.

Once these translation problems were outlined, León-Araúz & Faber (2014) listed nine different translation relations to tackle each one of these problems. The types of translation defined by the authors are:

- Canonical translations: applied when there is no translation problem, and the terms are symmetrical in terms of meaning.
- Generic-specific translations: applied to address problems 1, 2, and 3. It consists of the translations by means of hypernyms.
- Extensional translations: applied to address problems 1 and 2. It is a type of generic-specific translations because the original term is translated by all of

the hyponyms of the concept in the target culture, benefiting from the listing of its subtypes.

- Communicative translations: applied to address problem 4. This type of translation considers the communicative situation and can use an expert neutral variant as an equivalent, for example, if it is necessary. This type of translation establishes a register correspondence among domain-specific and diaphasic variants.
- Functional translations: applied to address problems 5, 6, and 7. They consist of deculturalising original terms so the receivers can relate to the concept. The chosen equivalent is the closest concept from a semantic point of view, but the cultural traits are lost.
- Cultural translations: applied to address problems 6, 7, and 8. Their aim is to avoid the translation impairment cause by cross-cultural differences, and they consist of adapting original culture-bound terms to other culture-bound terms in the target culture.
- Descriptive translations: applied to address problems 2, 7, 8, and 9. It consists of the addition of an extra word to explain the concept.
- Non-translations: applied to address problems 7 and 9. The original can be used if the receivers are experts in the field.
- Metonymic translations: applied to address problem 10. It consists of employing original lexical units in terms of their metonymic variant and leaving the target term in its original form.

Considering the theoretical background defined by us on chapter three of this work and highlighted briefly in this section, we must now consider the lexical equivalents which will be analyzed under this academic light. We will follow the same order of analysis adopted in section 5.1. Therefore, we will begin with the analysis of the equivalents for “plant” and “site”. Then we proceed to analyze the lexical equivalents for “material” and “article”. The next subsection begins the analysis.

5.2.1 Lexical equivalents of plant and site

We shall commence by listing all of the terms considered possible lexical equivalents for this pair of alignment terms. The following terms have been identified: “*fábrica”, “departamento”, “loja”, “loja física”, ‘loja âncora”, “varejo”, “unidade*

industrial", "unidade de negócio", "canal de distribuição", "fonte de suprimentos", and "*estoque*". "*Fábrica*" is the closest term to "factory" and is defined in Portuguese as an industrial establishment where raw materials are transformed into products. The definition can be found below:

- *Estabelecimento industrial onde se transforma matéria-prima em produtos. Exemplo: "A fábrica de massas italianas, ali mesmo da vizinhança, começou a trabalhar, engrossando o barulho com o seu arfar monótono de máquina a vapor".* [Industrial establishment where raw materials are transformed into products. Example: "The Italian pasta factory, right next door, started working, adding to the noise with the monotonous panting of a steam engine".]

In this case, it seems like "*fábrica*" is very much aligned with "factory", which was considered by us as a hypernym of "plant" and "site". Thus, if we compare this definition in Portuguese to our definitions of "plant" and "site" we can see a similarity. However, despite some shared semantic features, we cannot affirm that "*fábrica*" is a lexical equivalence for "plant" nor "site". Although, it could be used to refer to a specific type of site or plant in which goods are produced, due to the closeness of this meaning to the retail domain.

Next, we have the term "*departamento*", which could be translated as "department" and has the following definitions:

- *Repartição em alguma organização pública ou privada.* [Sharing of a public or private organization].
- *Subdivisão de uma organização administrativa.* [Subdivision of an administrative organization].

If we refer back to the definitions of "plant" and "site" posed by us (see Table 11), these terms can refer to parts of an enterprise. When it comes to "plant", it can be defined as "[…] the nodes in a hierarchy containing further plants […]". Meanwhile, the definition of "site" that is closest to this definition of "*departamento*" is worded as follows: "It is a separate, smaller facility at a plant where you manufacture a product, a specialized portion of a plant or facility, etc.". Hence, if we consider these definitions that specify a division in the site or plant, we could look at "*departamento*" as a possible lexical equivalent for these terms in Portuguese. But there are other factors to take into consideration.

Despite these semantic features, the fact that “plant” and “site” can refer to divided, more specific areas in a location is actually not the norm. Thus, we cannot classify “*departamento*” as an equivalent for these terms, although it could be used to refer to particular areas within a location. The next selected term to be analyzed is “*loja*”, which refers to a store. The definition for “*loja*” that matches this context is the one below:

- *Estabelecimento comercial onde se vendem mercadorias diversas ou um único produto. Exemplo: “Depois da refeição, Jango disse que queria comprar presentes para os filhos e me convidou para ir com ele. Como não podia ir no carro dele, tomei um táxi e o segui até a loja”.* [Commercial establishment in which different goods or a single product are sold. Example: “After the meal, Jango said he wanted to buy gifts for his children and invited me to go with him. Since I couldn't go in his car, I took a taxi and followed him to the store”.]

As we can draw from the definition, this lexical unit refers to a location concerned with the very specific commercial transaction that can happen in a plant or a site. When the article or material reaches the final stages of production, it is ready to be sold at a plant or site (or part of a plant or site) that is concerned with this type of activity in the retail sector. This is what “*loja*” refers to. Therefore, because it is so specific to the sales department, “*loja*” was not classified as a lexical equivalent for “plant” or “site”.

The same goes to the other two terms to be analyzed, which are “*loja física*” and “*loja âncora*”. They refer, respectively, to a physical store, as the contrary to an online store, and to a type of store that is a main store dedicated to attracting customers to the other stores. There is also the concept of a virtual store, but since one of the main characteristics of plant and site is the fact that they refer to physical locations, we excluded this one. The definitions of “*loja física*” and “*loja âncora*”, respectively, can be found below:

- *Estabelecimento de comércio com estrutura física. A loja física tem como principal característica elevar os custos iniciais de um projeto, momento que pode ser sensível financeiramente para quem está abrindo o próprio negócio. Por outro lado, um bom ponto comercial tem o papel importante de fazer com que as pessoas saibam que a loja existe*. [Commercial

establishment with a physical structure. The physical store's main characteristic is to raise the initial costs of a project, a moment that can be financially sensitive for those who are opening their own business. On the other hand, a good commercial point has the important role of letting people know that the store exists.]

- *Espécie de loja principal, uma loja de apoio que atrai os clientes para as demais. São as grandes lojas de um shopping center. Por regra, são áreas que ocupam mais do que 1.000m² dentro do empreendimento. Uma loja âncora é uma loja no shopping center (centro comercial) que tem um tamanho maior em relação as lojas comuns (lojas satélites) e que poderá atrair um público maior ao shopping, pois é conhecida nacionalmente ou internacionalmente. Quando um shopping traz para ele diversas lojas âncoras, o público do shopping e o consumo pode aumentar muito, ainda mais quando aquela loja não existia na região*. [Type of main store that is dedicated to attracting customers to the other stores. They are the big stores in a shopping mall. As a rule, these are areas that occupy more than 1,000m² within the mall. An anchor store is a store in the shopping center (commercial center) that has a larger size compared to regular stores (satellite stores) and that can attract a larger audience to the mall, as it is known nationally or internationally. When a mall brings several anchor stores to itself, the mall's audience and consumption can increase a lot, even more so when that store didn't exist in the region.]

In a similar way to "*loja*", "*loja física*" and "*loja âncora*" were not considered lexical equivalents for plant or site, especially because they refer to even more specific concepts than "*loja*" does. Now we move on to "*varejo*", which could be translated into "retail". The definitions are specified below:

- *Comércio no qual se vendem as mercadorias por unidade, por quilograma ou fração deste, exercido por revendedores – os varejistas – que adquirem os bens dos produtores ou dos atacadistas.* [Trade in which goods are sold by unit, by kilogram or fraction thereof, exercised by resellers – retailers – who purchase goods from producers or wholesalers.]
- *COMERCIAL - Maneira de vender certas mercadorias, diretamente ao consumidor final, sem passar por intermediários.* [COMMERCIAL - A way

of selling certain goods directly to the final consumer, without going through intermediaries.]

Once again, these definitions are very specific to the business side of the retail industry. They focus on the final stages of product distribution, which is concerned with the selling and buying of goods. Thus, "*varejo*" is not our best option for lexical equivalence of plant and site. "*Unidade industrial*" and "*unidade de negócio*", on the other hand, seem to offer a better alternative. We shall begin by analyzing the definition of "*unidade industrial*", outlined below:

- *É usual, no setor de engenharia, o uso da expressão planta industrial e muitas vezes apenas planta, com o significado de uma unidade industrial, ou mesmo setor dentro de uma indústria, que realize algum processo específico. Exemplos: "planta de fundição, planta de moldagem, planta de extrusão, planta de estamparia, planta de laminação".* [It is usual, in the engineering sector, to use the expression industrial plant and often just plant, with the meaning of an industrial unit, or even sector within an industry, that performs some specific process. Examples: "foundry plant, molding plant, extrusion plant, stamping plant, rolling plant".]

It is worth noting that we were not able to find a definition of "*planta*", which would be a literal translation of "plant", regarding the retail domain. In this definition of "*unidade industrial*", however, we can observe that one can refer to an industrial unit as "*planta industrial*" in Portuguese. This leads us to believe that the use of "*planta industrial*" to refer to a "*unidade industrial*" is so domain specific that it cannot be found online or even in sources concerned with the domain. Additionally, the fact that it is a lexical unit formed by two words, instead of just one, makes it a challenge to be looked up in dictionaries or machine translation tools. In this case, the definition of "*unidade industrial*" (or "*planta industrial*", as the definition clarifies) seems to be applicable to many sectors of a plant or site or even to the totality of the location. The activities described in the definition are not detailed to the point of exhaustion, which leaves the door open for interpretation. As the examples clarify, this terminology can be used to describe a variety of establishments, working as a hypernym for more specific parts of a plant or site. Due to these characteristics, we considered both "*unidade industrial*" and "*planta industrial*" as lexical equivalents of "plant" and "site".

Next, we must consider "*unidade de negócio*". This lexical unit was proven to be a rich source of definitions, and we were able to find four of them. We will include two of them here with the purpose of omitting repetitive information due to the length of these definitions.

- *Uma unidade de negócios é uma divisão, linha de produtos ou outro centro de lucro de uma empresa que produz e comercializa um conjunto bem definido de produtos ou serviços correlatos, serve um conjunto claramente definido de clientes, numa área geográfica razoavelmente bem delimitada e compete com um conjunto bem definido de concorrentes.* [A business unit is a division, product line, or other profit center of an enterprise that produces and markets a well-defined set of related products or services, serves a clearly defined set of customers, in a reasonably well-defined geographic area, and competes with a well-defined set of competitors.]
- *Uma unidade de negócio de uma empresa é aquela que opera de forma independente, portanto, tem missão e objetivos próprios; o que permite que o seu planejamento seja realizado de forma autônoma em relação às demais unidades da empresa. Claro, a separação por unidades de negócios é feita especialmente em empresas que são muito grandes e que produzem muitos produtos diferentes ou têm como alvo diferentes grupos de mercados. O conceito de unidade de negócio permite extrair resultados a partir de uma análise de lucratividade segmentada.* [The business unit of a company is one that operates independently, therefore, it has its own mission and objectives, which allows its planning to be carried out autonomously in relation to the other units of the company. Of course, the separation by business units is especially done in companies that are very large and that produce many different products or target different groups of markets. The business unit concept allows the result extraction from a segmented profitability analysis.]

Once again, we find definitions which are very specific to types of plant and types of site. To consider "*unidade de negócio*" an equivalent of plant or site would be to narrow its meaning down to a smaller fraction of what the concept represents. Thus, this term is not considered a lexical equivalent in this case. Next, we must address

the term "*canal de distribuição*", which is also an option we have. The definition for this term can be found below:

- *Canais de distribuição são os meios pelos quais uma empresa fabricante escolhe entregar seus produtos ao consumidor final. Esses canais são como o comprador tem acesso aos bens. Toda empresa que trabalha no ramo comercial tem um sistema interativo por meio do qual entrega seus produtos ao consumidor final. Esse processo é conhecido como canal de distribuição e, nele, participam o fabricante, o meio que o distribui e, por fim, o consumidor.* [Distribution channels are the means by which a manufacturing company chooses to deliver its products to the final consumer. These channels are how the buyer gets access to the goods. Every company that works in the commercial field has an interactive system through which it delivers its products to the final consumer. This process is known as the distribution channel and, in it, the manufacturer, the means that distribute it and, finally, the consumer participate.]

As we can observe, this term does not work as a lexical equivalent for neither "plant" nor "site". Despite being related to the retail industry and the commercial transaction sphere, this term does not acknowledge the other activities which could take place in a site or a plant. Another related lexical unit is "*fonte de suprimentos*", which could be translated as "supply source". The definition is stated below:

- *Fornecedor de material encomendado de qualquer tipo de organização*. [Supplier of ordered material from any type of organization.]

This definition focuses mainly on the organization's supplier and does not take into consideration a specific location, or any other activities related to supplying something to someone. Thus, it is clearly not a case of lexical equivalence for "plant" or "site". Lastly, we shall analyze "*estoque*", which is the remaining candidate to lexical equivalence that is yet to be analyze. Its definition in the retail domain is stated as follows:

- *Local (depósito, armazém, silo etc.) onde essa mercadoria é guardada. Exemplo: "No estoque só entram os funcionários autorizados."* [Location (warehouse, silo, etc.) where this merchandise is kept. Example: "Only authorized employees enter the stock."]

In this case, this definition matches the following quote from our own definition of plant, in which we describe the term as an organization unit for dividing an enterprise according to the sectors related to handling goods, from the very assembly of a product to its stocking and distribution. However, this term, despite being different from the other terminology analyzed so far, lacks in the same way. It focuses mainly on one type of activity related to sites or plants, instead of contemplating the whole. Therefore, "*estoque*" cannot be classified as a lexical equivalent for "plant" nor "site".

In terms of lexical equivalents for the terminology at hand, we have identified two terms that could be plausible equivalents for "plant" and "site" in Portuguese. These terms are "*unidade industrial*" and "*planta industrial*". The first one was analyzed here by us, and it appeared in the specialized dictionaries we consulted. The second one is so domain-specific that we could not find it in our corpus or in these thesauruses. As previously mentioned, we also believe that the fact that this translation requires an extra word to fully contemplate the sense of "plant" makes it an extra challenge when we are trying to translate "plant" into a word in Portuguese. It is a case of structural divergence, as described by L'Homme (2020).

If we try to classify the candidate terms for lexical equivalence in this case according to the translation problems they represent, taking the definitions of León-Araúz & Faber (2014) as guidance, we could rule out problems number 2, 4, 7, 8, and 9. Problems number 7 and 9 do not concern these terms because the entities referred to (plant and site) exist in both cultures we are handling. Problem number 2 can be ruled out because there are terms in both cultures to refer to these entities. Problem number 4 can be disregarded because the concept is lexicalized in both cultures, not just in one. And problem number 8 can be disregarded due to the fact that the terms have not been recycled from one language to the other to refer to different concepts.

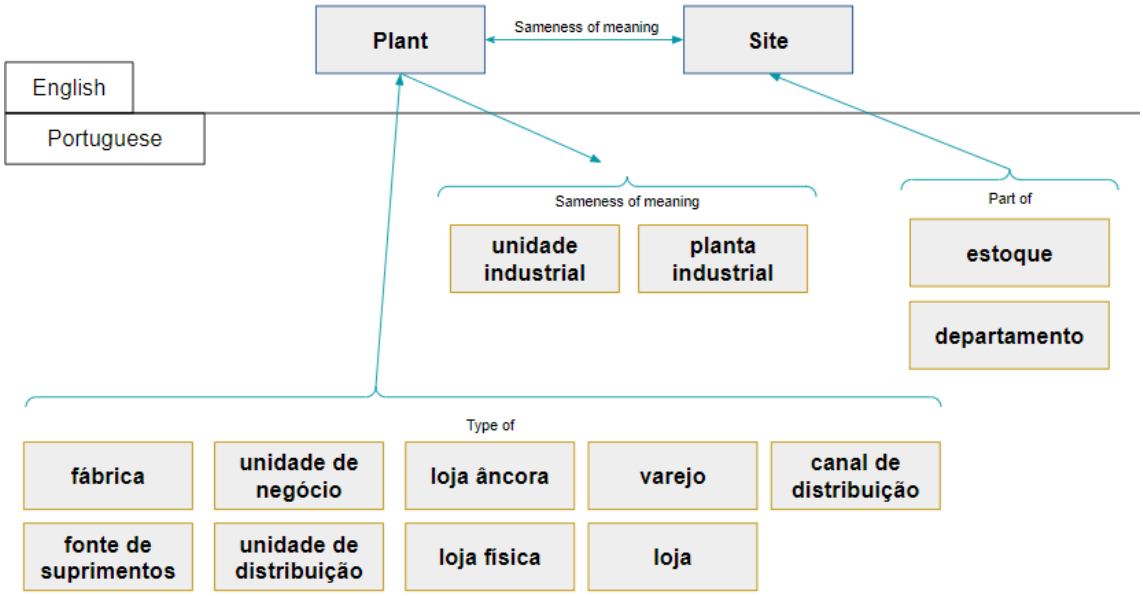
Hence, we are left with problems 1, 3, 5, 6, and 10 to contemplate. We believe we can ignore problem number 6 because the adopted terminology comes from the specialized domain and the translation challenges are not rooted in cultural differences or asymmetries. Problem number 5 can be dismissed as well because the issue in this case does not regard design principles or different structures. Since "*planta industrial*" and "*unidade industrial*" were classified as equivalents, they do not pose any translation problems and will not be considered in this classification.

We believe that the terms “*departamento*” and “*estoque*” refer to parts of a plant or site, establishing a metonymic designation between languages. Thus, these terms have been considered to be related to problem number 10. The remaining terms, which are “*fábrica*”, “*loja*”, “*loja física*”, “*loja âncora*”, “*varejo*”, “*unidade de negócio*”, “*unidade de distribuição*”, “*canal de distribuição*”, and “*fonte de suprimentos*” could be related to both problems 1 and 3 since it could be affirmed that the terms vary from more specific or more general than the source terms and they also seem to highlight different aspects of the concept in each language.

In terms of the translation relations described by León-Araúz & Faber (2014), we have identified the seventh type, the descriptive translation as the type of translation applied to the terms “*planta industrial*” and “*unidade industrial*”. “Plant” and “site” are terms that do not require any extra words to complete their senses in the retail domain. However, “*planta*” needs the addition of “*industrial*” just like “*unidade*” does because otherwise, their senses would not be related to the domain. These translations explain the particular function of a plant or site. Thus, they are classified as descriptive.

In order to close this subsection of the work, we have elaborated a conceptual map considering the terms “plant” and “site”, and their relations with the terms in Portuguese. Figure 34 represents this visualization.

Figure 34 – Conceptual map of the cross-lingual relations for plant and site



Source: made by the author (2023).

These considerations close our analysis of the lexical equivalents for “plant” and “site”. In the next subsection, we will evaluate the chosen equivalents for the other pair of terms analyzed by us, which is “material” and “article”. The structure of subsection 5.2.2 follows the same order and organization of this subsection.

5.2.2 Lexical equivalents of material and article

In this subsection, we intend to follow the same analytical path adopted by us in the first subsection, only this time our focus shifts to the pair of terms “material” and “article”. The identified possible lexical equivalents for this pair are: “*artigo*”, “*produto*”, “*objeto*”, “*unidade*”, “*mercadoria*”, “*bem intermediário*”, “*bem de comparação*”, “*bens de conveniência*”, “*commodity*”, “*estoque*”, and “*granel*”. We shall begin by closely examining the definitions of “*artigo*”, “*produto*”, “*objeto*”, and “*mercadoria*”. The reason why we decided to analyze this set of terms is because they are closely related in terms of semantic information. Their descriptions can be found below:

- *ARTIGO - Objeto posto à venda; mercadoria.* [Object offered for sale; merchandise].
- *PRODUTO - O que é produzido, destinado ao consumo próprio ou ao comércio. Exemplo: “Ele pediu para que eu lesse o rótulo do produto e identificasse os ingredientes.”* [What is produced, intended for own consumption or trade. Example: “He asked me to read the product label and identify the ingredients].
- *PRODUTO - Conjunto de bens e serviços que resultam da atividade produtiva de uma nação, de uma empresa ou de um indivíduo*. [Set of goods and services that result from the productive activity of a nation, a company or an individual].
- *OBJETO - Qualquer coisa a ser comercializada; artigo, mercadoria.* [Anything to be marketed, article, commodity].
- *MERCADORIA - Qualquer bem que pode ser comprado ou vendido*. [Any asset that can be bought or sold].

The first lexical unit in Portuguese suffers of the same problem as its counterpart in English: most of its senses are related to a piece of writing with a well-defined

structure. However, if we compare this definition to the definition of article elaborated by us (see Table 11), it is possible to associate it with most of the definition. “Article” is described as “an object subject to commercial transactions that includes products, materials, articles, and services”. “Material” was described in a similar way, representing “an object that includes products, materials, articles, and services and that is the subject of business activity”. Thus, based on these shared semantic traits, we classified “*artigo*” as an equivalent for “article” and “material”.

The next lexical unit we must analyze is “*produto*”, which could be translated as “product”. In this case, we have identified two definitions which would fit the context and be related to the retail domain. It is possible to notice that “*produto*” does describe a physical object destined to commercialization and it shares some semantic information with “article” and “material”. It could be said that “*produto*” works as an equivalent, but in this case, the specialized terminology comes into the picture. Although “*produto*” could be a lexical equivalent used to describe “article” and/or “material”, it lacks the conceptualization provided by the domain and it does not constitute specialized terminology. Thus, since we are considering a specialized domain when addressing the lexical variants and also the lexical equivalents, we opted for the non-classification of “*produto*” as an equivalent.

The same goes for “object” which closely aligns with the definitions of the terminology under analysis. However, just like it happened to other terms, the sense is too broad to account for the specificities of the domain. Additionally, “*objeto*” itself is not part of a specialized domain terminology. Therefore, it was not classified as an equivalent for “material” and “article”. “*Mercadoria*” has a similar definition and poses a similar challenge to translation. Thus, for the reasons stated above, it was also not classified as an equivalent.

“*Unidade*” and “*granel*” are the terms we shall consider next. They are also being analyzed together due to the similarities of meanings. These similarities can be observed in the definitions below:

- *UNIDADE - Cada item de um conjunto de objetos produzidos em série, considerado individualmente. Exemplo: “A fábrica tem capacidade para produzir 100 mil unidades de computadores por ano.”* [Each item from a set of objects produced in series, considered individually. Example: “The factory has the capacity to produce 100,000 units of computers per year].

- *GRANEL - Mercadoria comercializada fora da embalagem, em quantidades fracionárias.* [Merchandise sold unpackaged and in fractional quantities].

The literal translation of "*unidade*" would be "unit" in English and just like "unit", we observe that "*unidade*" can be used to refer to a specific item from a set of items produced or sold. Despite the fact that our definitions of "material" and "article" include references to a single unit of a product or the smallest unit in a costumer pack, this term, in Portuguese, can be used to refer to a variety of objects. We would feel much more comfortable if "unit" were considered a valid term for alignment in this domain. But since it is not, we will not consider "*unidade*" an equivalent due to the fact that it lacks the status of specialized terminology, just like "*objeto*" did. The same goes for "*granel*". There are semantic approximations, but the fact that it is not part of the specialized terminology holds more power when it comes to lexical equivalence, since it can alter the sense and the message in the target context. Next, we have three similar items: "*bem intermediário*", "*bem de comparação*", and "*bem de conveniência*". Their definitions are as follows:

- *BEM INTERMEDIÁRIO - Matéria-prima ou bem manufaturado processado que é empregado para a produção de outros bens ou produtos finais.* [Raw material or processed manufactured good that is used for the production of other goods or final products].
- *BEM DE COMPARAÇÃO – Artigo cuja compra é infrequente, dependendo de planificação e pesquisa de informação, comparação das marcas em qualidade, atributos, estilos e preços.* [Article whose purchase is infrequent, depending on planning and information research, comparison of brands in quality, attributes, styles, and prices].
- *BEM DE CONVENIÊNCIA - Artigo adquirido com frequência pelo consumidor, de imediato e com um esforço mínimo*. [Item purchased frequently by the consumer, immediately and with minimal effort].

In this case, we have found three different terms which refer to material and article, but they refer to different types and characteristics of these goods being sold. The first one resembles a type of material, from which one can produce the goods. Thus, it would hold a hyponymic relation with the original terms. The other two terms are used to differentiate between goods that are frequently sold and goods that are not commonly purchased. Since neither "article" nor "material" marks these

distinctions, they are not lexical equivalents. “*Bem intermediário*” is not a lexical equivalent as well because it is too specific to represent “material” or “article” in a satisfactory way. Next, we have “*commodity*”, which has several definitions:

- *Termo que designa substância física primária sujeita a escambo com outras do mesmo tipo, cujos investidores as compram e vendem, normalmente, por meio de contratos a termo – contratos futuros – em bolsas de mercadorias. Hoje também são considerados commodities produtos de uso comum mundial, como lotes de camisetas ou de calças jeans*. [A term that designates a primary physical substance subject to exchange with others of the same type, whose investors buy and sell them, normally, through forward contracts – futures contracts – on commodity exchanges. Today, products of common use worldwide are also considered commodities, such as batches of T-shirts or jeans].
- *Mercadoria em estado bruto ou produto básico de grande importância no comércio internacional, como café, cereais, algodão etc., cujo preço é controlado por bolsas internacionais*. [Raw merchandise or basic product of great importance in international trade, such as coffee, cereals, cotton, etc., whose price is controlled by international exchanges].
- *Qualquer produto em estado bruto relativo à agropecuária ou à extração mineral ou vegetal, de produção em larga escala mundial, dirigido para o comércio internacional*. [Any raw product related to agriculture, livestock or mineral or vegetable extraction, produced on a large scale worldwide, intended for international trade].

“*Commodity”* is an interesting case since it is a term borrowed from English and broadly used in Brazilian Portuguese to refer to goods. It has been growing in popularity and it is widely used in the retail industry. As we can observe, it can be used to describe a variety of articles and materials, from raw products to fully assembled and manufactured goods. For this reason, it was considered a lexical equivalent for “material” and “article”.

Lastly, we turn our attention to “*estoque*”. Its definition is stated as follows:

- *Quantidade de mercadoria armazenada de que se dispõe para uso, venda, doação, exportação etc.: “O estoque de sorvete esgotou naquele dia.”*

> [Amount of stored merchandise available for use, sale, donation, export, etc.: “The ice cream stock ran out that day.”]

In this case, the lexical unit refers to a set of articles and materials, or huge quantities of goods. Since none of the definitions identified for “material” and “article” (see Table 11) refer to the quantity of these elements (although they do mention the fact that these terms could refer to more than one unit), we decided not to classify “*estoque*” as a lexical equivalent in this case.

To summarize this part of the analysis, we can affirm that there were two lexical equivalents identified for “material and “article” in Portuguese. These equivalents are “*artigo*” and “*commodity*”. The first one is a literal translation of “article” and the second one is a term adopted from English and widely used in Portuguese when it comes to retail. We believe that “commodity” was probably adopted due to its meaning in English, which is similar to the meaning in has in Portuguese.

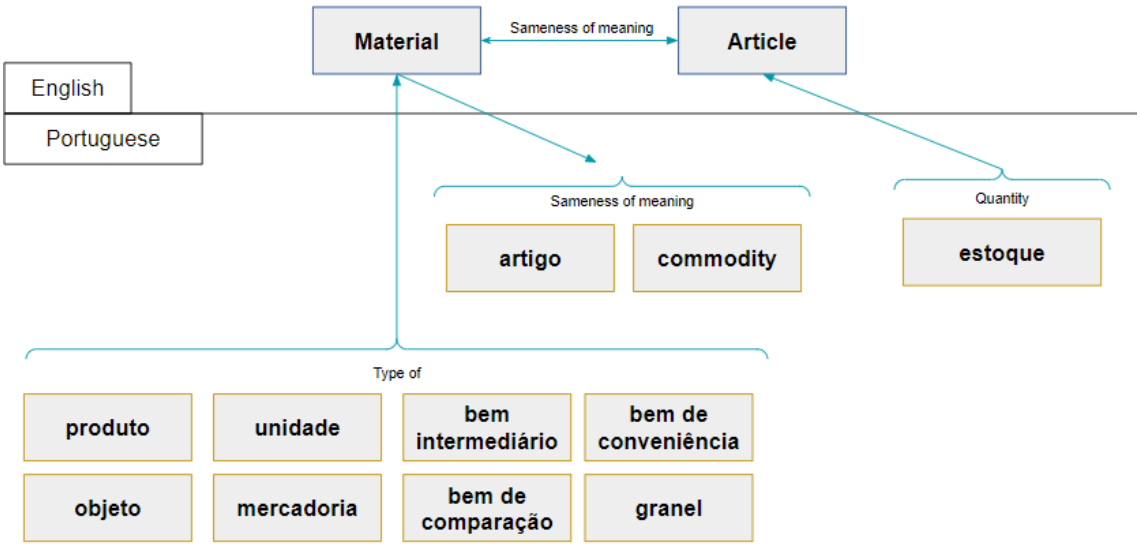
If we try to classify the candidate terms for lexical equivalence in this case according to the translation problems they represent, taking once again the definitions of León-Araúz & Faber (2014) as guidance, we could rule out problems number 2, 4, 5, 6, 7, 8, 9, and 10. Problems number 7 and 9 do not concern these terms because the entities referred to (material and article) exist in both cultures we are handling. Problem number 2 can be ruled out because there are terms in both cultures to refer to these entities. Problem number 4 can be disregarded because the concept is lexicalized in both cultures, not just in one. Problem number 5 can be dismissed as well because the issue in this case does not regard design principles or different structures.

We believe we can ignore problem number 6 because the adopted terminology comes from the specialized domain and the translation challenges are not rooted in cultural differences or asymmetries. Problem number 8 can be disregarded due to the fact that the terms have not been recycled from one language to the other to refer to different concepts. And problem number 10 can be ignored in this case because the terms do not refer to parts of a material or article and do not establish a metonymic relation between languages. Because “*artigo*” and “*commodity*” were classified as equivalents, they do not pose any translation problems and will not be considered in this classification.

Hence, we are left with problems 1 and 3 to contemplate. We believe that the terms "*produto*", "*objeto*", "*unidade*", "*mercadoria*", "*estoque*", and "*granel*" contemplate problem number 1, since they are more generic terms in some cases or denote more specific aspects of the concept in other cases. The terms "*bem intermediário*", "*bem de comparação*", and "*bem de conveniência*" associate with problem number 3 since they highlight different aspects of the concept and focus on it from a different perspective, the perspective of the type of product and how much it sells.

In terms of the translation relations described by León-Araúz & Faber (2014), we have identified the first type, the canonical translation as the type of translation applied to "*artig*o" and the eighth type, the non-translation for the term "*commodity*". "*Artigo*" was classified this way because in this context, there seem to be no translation problems when looking for an equivalent. "Commodity", on the other hand, was classified as a non-translation because the term was originally an English word, and it was adopted by Portuguese speakers to refer to the concept. Although the entity exists in Portuguese, we still consider this a non-translation because the lexical unit is kept in its original form. Lastly, we have elaborated a conceptual map which works as a visual representation of the interactions between the terms in Portuguese and English. These lexical units in Portuguese can be considered variants of the "*artigo*" and "*commodity*" equivalents. Figure 35 represents the map.

Figure 35 – Conceptual map of the cross-lingual relations for material and article



Source: made by the author (2023).

With these considerations, we close our analysis chapter having conducted a monolingual and a bilingual analysis. The monolingual analysis took into consideration the predictions made by three lexical substitution models executed by us and the theoretical postulates made by Apresjan (1974), Murphy (2003, 2010), Croft & Cruse (2004), and León-Araúz & Faber (2014). The bilingual analysis, on the other hand, took into account the equivalents elected by us and the academic considerations posed by León-Araúz & Faber (2014). Our next chapter outlines the final considerations of this work and also contributes to this analysis by summarizing our thoughts and providing future contributions to this study.

6 FINAL CONSIDERATIONS

The main objective of this work was to investigate the phenomenon of lexical variation in Portuguese and English in the term alignment and lexical substitution stages in NLP. The secondary objectives were to investigate to what extent the analysis of these elements could be helpful to the process of improving NLP models aimed at lexical substitution, and to analyze what types of lexical variation occur during the lexical substitution process and find out to what extent the multilingual aspect is affected by such lexical variants. In order to accomplish these goals, we relied on the theoretical background provided by the Computational Lexical Semantics field and the precepts of Terminology and Lexical Semantics aimed at Terminology. As for the analysis, we focused on the Retail domain terminology to narrow our investigations. Therefore, our work was divided into the following chapters.

Chapter two, titled “Semantics in Computational Linguistics”, intended to dive into the Natural Language Processing field of study, since our work is closely related to it, and approach a more specific field within it: the Computational Lexical Semantics field. We approached the state of the art regarding studies focusing on Semantics and NLP along with the challenges related to processing natural language considering semantics. On the second part of the chapter, we focused on computational lexicons that aimed at representing semantics and semantic relations in the online environment and were successful in accomplishing such bold goal.

However, in order to fully contemplate the theoretical scope of our work, it was necessary to approach the semantic relations and the terminology studies of lexical variants with the lens of Linguistics. Thus, chapter three focused on these aspects of our work. We discussed the role of lexical semantics in Terminology considering both a monolingual and a bilingual approach. Since we focused on semantics and NLP in chapter two, we decided to give semantics in terminology – monolingual and bilingual – a more salient space in this chapter. Therefore, chapters two and three complement each other because they provide insights on Semantics and Terminology from different perspectives: the Computational one, in chapter two, and the Linguistic one, in chapter three.

Chapter four described the methodological steps adopted by us in order to accomplish the objectives we had in hands. We began by describing the materials to

be used, which were the tool called Sketch Engine and the methodological guidelines provided by Corpus Linguistics. Moreover, we described the models used to gather the term-related data and the steps adopted in the semantic analysis we intended to perform. Therefore, we were able to cover both the materials and the methods used.

Chapter five focused on the analyses of monolingual and bilingual terminology related to the retail domain. We outlined the analyses made following the methodological steps described in chapter four and supported by the theories and postulates presented in chapters two and three. To accomplish the goals of this study, we used a semantic-terminological approach to examine and explore the data. This gave us the opportunity to classify the terminological variation according to the categorization of León-Araúz & Faber (2014) and the terms considering the semantic relations explained by Apresjan (1974), Murphy (2003, 2010), L'Homme (2020), and Croft & Cruse (2004) in the monolingual stage. When it comes to the bilingual analysis, we were able to identify the correct equivalents in Brazilian Portuguese for the terminology used in the retail domain and classify this data according to the translation problems and translation relations defined by León-Araúz & Faber (2014).

Our conclusions regarding the analyses made were based not only on observations of the results, but also taking into account other studies developed in the area, which focus on similar issues. Firstly, when it comes to the monolingual analysis, we could observe that all the models were able to disambiguate at least one pair of terms. "Plant" and "site", for example, are polysemous terms. "Plant" can convey not only the retail domain related meaning, but also a botanic meaning, described as "a living organism of the kind exemplified by trees, shrubs, herbs, grasses, ferns, and mosses, typically growing in a permanent site, absorbing water and inorganic substances through its roots, and synthesizing nutrients in its leaves by photosynthesis using the green pigment chlorophyll". Moreover, "site" also expresses a computational sense, meaning a website. All the models could overcome these hurdles posed by the existence of more than one sense linked to both terms.

However, the same did not happen with "material" and "article" and we believe that the hypernym nature of these terms may have had an effect on the outcomes. In this case, more information is needed in order for the models to be able to disambiguate the terms and suggest better lexical substitution predictions. There is also the fact that the models were operating on different sets of linguistic information. Moreover, one could argue that four terms are not enough to classify these models

as adequate or not, but our analyses have considered a detailed degree of investigation. It contributes to the sense that it opens doors for the analysis of a bigger number of terms, contributing to more results. Additionally, it is relevant to mention that there were more analyses executed by the VLHSem Group that corroborate to this conclusion yet were not mentioned here because they belong outside the scope of this work.

One hypothesis raised by us in chapter five was that the synonymy information, the word embeddings, and the semantic frames are not enough (in the cases of the four terms in our data) information for the models. Yet, when we consider these models in the scope of the VLHSem Project, it is possible to observe improvements, especially when the models learn how to operate by considering the semantic frames. The fact that we have more data available helps us to reach this conclusion. However, there is no reason why more linguistic information should not be added to the lexical substitution models. Using the term variants classification posed by León-Araúz & Faber (2014) could be an option. Linguists could operate by compiling a list of domain-related terminology and classifying each term according to the variants listed by León-Araúz & Faber (2014). Then, the models could be trained using this information and the semantic frames. This combination could culminate in more accurate suggestions, given that the model would have access to domain-related terminology, since this is the type of variant that appeared the most in our monolingual analysis. This is a hypothesis that needs to be investigated further on.

One could also argue that hyponymy and meronymy pose a more challenging environment for the models than synonymy does, since these were the majority of the relations concerning “material” and “article”. There is, however, a considerable amount of synonymy relations, especially when we compare “article” with the adequate choices for lexical substitution chosen for this term. Thus, before training the lexical substitution models to be sensitive to a latent degree of hyponymy and meronymy, there is the need for further analyses regarding this point. The combination of these terminological and semantic aspects of terminology could help us improve our models even more.

One point that cannot be argued, though, is the usability of Frame Semantics in the models developed by us. As mentioned before, when analyzing huge sets of terms, it is possible to observe improvements in the results. The models are able to provide more accurate terminology when they consider the frames. Our intend further

on is to develop more domain-related frames, since according to L'Homme (2020), Frame Semantics is helpful to Terminology because it accounts for the participants involved in the meaning of terms and the behavior of terms. But the impact of Frame Semantics in terminological studies does no stop here. It also contributes to the connection of linguistic properties of terms to conceptual representations, such as the frames. Thus, it offers a unique way to capture complex semantic phenomena (L'HOMME, 2020).

When it comes to the bilingual analysis, we were able to find a case of what L'Homme (2020) calls structural divergence. The equivalents of "plant" and "site" in Portuguese are "*unidade industrial*" and "*planta industrial*", which can be regarded as cases of structural divergences since they are lexical units formed by more than one word. This can pose more problems to the translator because we are considering a specialized field. Doors are now opened to mismatches in translation and confusion, especially if the translator is not familiar with the field at hand. If the translator is a system, there is, if we are dealing with a case of machine translation, then the program must be sensitive to these issues and able to reach the terms "*unidade industrial*" and "*planta industrial*" instead of "*planta*" and "website", which would be the literal translations of "plant" and "site" in Brazilian Portuguese. It would require a machine translation program that considers descriptive translations.

As for the equivalents of "material" and "article", there are two possibilities as well: "*artigo*" and "commodity". The first one was considered by us a canonical translation since it does not differ from a literal translation and does not pose many challenges. The second one, on the other hand, consists of a borrowing of an English word. In this case, it was classified in our analysis as non-translation, since the word was added to the Brazilian Portuguese vocabulary with no alterations regarding spelling, pronunciation, or meaning. This reveals to us a need of machine translations that consider these different types of translation elicited by León-Araúz & Faber (2014), since there are multiple types of equivalents that can be found in specialized translation.

As stated by Boas (2009), one of the main problems of creating multilingual lexical databases is the development of a system capable of managing a wide range of linguistic challenges such as diverging polysemy, differences in lexicalization patterns, and translation equivalence. If we were trying to create a multilingual lexical database taking into account only "plant", "site", and their equivalents, we would

already run into a problem described by L'Homme (2020) and addressed by us in chapter three, which regards conceptual and lexical equivalence. Both "plant" and "site" can lead us to other senses since they are polysemic terms. If either one of these terms lead us to "factory" or "department" when organizing the information according to similarity of senses, there are multiple subtypes of factories and departments that could be listed. The same could happen in Portuguese if "*unidade industrial*" and "*planta industrial*" would also lead us to other terms and so on.

This confusion could be softened by taking Frame Semantics into account. The semantic frames would work as an organizing principle to help disambiguate word senses and organize both terms and their equivalents with respect to the domain. Of course, this would require the development of specialized semantic frames that account for the specialized background knowledge. FrameNet and its counterparts in other languages, although being general language enterprises, have already proven that efforts can be made in this direction.

Thus, we can list a few possibilities of future studies to be developed in the scope we work with, including the development of lexical substitution models in NLP that are able to identify domain-related terminology by using frame semantics and lexical semantics (considering the relations studied here, which are synonymy, hyponymy, and meronymy). There is also the possibility of investigating the development of machine translation models in NLP that are adapted to consider translation problems and the translation equivalents that aim to ease these issues, and the development of multilingual and cross-lingual language models that consider structural divergences in specialized language terminology and that consider descriptive translations, non-translations, among other equivalence problems that affect the lexicon and conceptualization across languages. Moreover, the development of domain-related semantic frames in order to map very specific conceptualizations in one language or more could also be a possibility.

Considering what has been mentioned so far, we believe that we were able to accomplish our main goal of investigating lexical variation in Portuguese and English in the term alignment and lexical substitution stages in NLP. Despite analyzing two pairs of terms, which could be deemed a small number of terms to consider, we were able to deeply examine how they behave semantically and terminologically when in relation to other terms. We were also able to find their equivalents in specialized language and observe what these equivalents imply to the translation stage. These

observations allowed us to offer insights on what challenges these pairs could pose for NLP systems aimed at lexical substitution and machine translation.

When it comes to the secondary objectives, we believe they were accomplished as well. The first secondary objective was to investigate to what extent the analysis of these elements can be helpful to the process of improving NLP models aimed at lexical substitution. As stated before, our classifications elicited linguistic aspects of the terminology that could be taken into account by NLP systems, depending on their demands and objectives. The second secondary objective was to analyze what types of lexical variation occur during the lexical substitution process and find out to what extent the multilingual aspect is affected by such lexical variants. We were able to not only identify the variant types that occur but also their semantic relations when interacting with the target terms. These findings, aligned with the theoretical background chosen, allowed us to link the variant types with equivalent problems that could arise from these variants and the equivalent types that could be used to ease these cross-lingual mismatches in specialized language.

As previously mentioned, this work has limitations. One of them is the amount of data analyzed. Despite having analyzed around 80 terms in total, it is considered a small number for computational models like Roberta to work with, for example. However, because these analyses were performed by humans and not models, 80 terms seem to be enough to reach the conclusions regarding semantics and equivalence that we presented in this study. We also believe that there is a gap in this scope of study and definitely more room for studies that link Linguistics and Computing, for Semantics has been little explored in this interface. There are also possibilities of study in the Linguistic field, if one is interested in researching the connections between Semantics, Terminology, and Translation Studies, since this is a relatively new interface in linguistic studies.

Despite the limitations, there are contributions. The relevance of this work lies in the fact that, as mentioned in the paragraph above, there is a gap when it comes to semantic studies in NLP approaches. Our intent was to provide some input to the linguists and computer engineers who are working with this very specific interface and looking for ways to improve their NLP systems with the help of linguistic information. There has been a rise in studies concerned with the usefulness of Cognitive Linguistics and Frame Semantics in technology development and we

believe this work contributes to this progress. Moreover, our work contributes to the demands posed by the VLHSem Project.

Speaking of the VLHSem Project, there are also contributions that concern the interdisciplinary character of this work. Our study combines two areas that have been growing as new technologies and ways to navigate the world appear. This interdisciplinary practice breaks with traditional patterns that prioritize the construction of knowledge in a fragmented way by viewing Applied Computing and Applied Linguistics as completely unconnected subjects. Instead, we reveal common points and favor critical analyses about different approaches to the same phenomenon by combining these subjects to accomplish one similar goal. Finally, this interdisciplinary interface is already proving itself to be relevant in current linguistic and technological areas since there are articles and abstracts produced by the VLHSem that have been submitted to journals and conferences, one of them already accepted[1].

Lastly, it can be said that our work contributes to the development of the Applied Linguistics field in Brazil. We have built this study in the scope of Applied Linguistics as it being a field concerned not only with theoretical development, but also with the usability of its theories, and an interdisciplinary nature. Thus, according to the definitions of Applied Linguistics and its concerns, we contributed to this context of work as well. There is also the fact that the Graduate Program of Applied Linguistics at UNISINOS University is a rich environment for research like this one and many others, related to Education, language teaching, conversation analysis, discourse analysis, among others.

Finally, we also believe that this study will be helpful in the future due to the rise of technology and the need for globalization. Linguistic information, and especially Cognitive Semantics and Terminology, has much more to contribute to these demands posed by society and the recent trends in the market. After all, language, in many environments, works as a bridge between cultures, people, and realities. Thus, adequately using language in NLP contributes to the structuring of a virtual space that reflects the reality we live in.

[1] The work titled “Integrating Frame Semantics in Lexical Substitution Tasks to Improve Lexical Precision” was accepted at the 16th International Cognitive Linguistics Conference on 02/22/23.

REFERENCES

AHRENBERG, L. **Alignment.** *In:* SIN-WAI, C. The Routledge Encyclopedia of Translation Technology. Routledge: New York, 2015.

AREFYEV, A. *et al.* **Always Keep your Target in Mind**: Studying Semantics and Improving Performance of Neural Lexical Substitution. COLING, 2020.

ALVES, I. M. **Polyset:** Modelo Linguístico-Computacional para a Estruturação de Redes de Polissemia de Nominais. Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopldo, 2009.

APRESJAN, J. **Regular Polysemy**. Linguistics, 1974, p. 5 – 32.

BAI, H. *et al*. **Better Language Model with Hypernym Class Prediction**. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1352–1362, Dublin, Ireland. Association for Computational Linguistics. 2022.

BAKER, M. **The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators**. International Journal of Corpus Linguistics, 1999.

BATES, M.; BOBROW, R. J. & WEISCHEDEL, R. M. **Critical Challenges for Natural Language Processing.** *In:* BATES, M. & WEISCHEDEL, R. M. Challenges in Natural Language Processing. Cambridge University Press, 2006.

BOAS, H. C. **From Theory to Practice**: Frame Semantics and the Design of FrameNet. Semantisches Wissen im Lexikon, Semantisches Wissen im Lexikon. Tübingen: Narr., 2005.

BOAS, H. C. **Semantic Frames as Interlingual Representations for Multilingual Lexical Databases.** In: Multilingual FrameNets in Computational Lexicography: Methods and Applications, ed. by H. C. Boas, 59–99. Berlin: Mouton de Gruyter. 2009.

BOAS, H. C. **Frame Semantics and translation**. *In:* A. Rojo and I. Ibarretxte-Antunano (eds.), Cognitive Linguistics and Translation. Berlin/New York: Mouton de Gruyter, 2013d, p. 125 - 158.

CABRÉ, M. T. **La terminología** - teoria, metodología, aplicaciones (trad. castelhana de Carles Tebé). Barcelona: Editorial Antártida/Empúries, 1993.

CHISHMAM, R. L. O. *et al.* **Dicionário FIELD**: dicionário de expressões do futebol. São Leopoldo: Unisinos, 2014. Available at: < http://dicionariofield.com.br/>. Accessed in: 10/09/22.

CHISHMAM, R. L. O. *et al.* **Dicionário Olímpico**. São Leopoldo: Unisinos, 2016. Available at: <http://dicionarioolimpico.com.br>. Accessed in: 10/09/22.

CHISHMAM, R. L. O. *et al.* **Dicionário Paralímpico**. São Leopoldo: Unisinos, 2021. Available at: < https://dicionarioparalimpico.com.br/>. Accessed in: 10/09/22.

CIMIANO, P.; MONTIEL-PONSODA, E.; BUITELAAR, P.; ESPINOZA, M. & PÉREZ, G. A. **A Note on Ontology Localization.** Applied Ontology, 2010.

CROFT, W., CRUSE, D. A. **Cognitive Linguistics**. Cambridge: Cambridge University Press. 2004.

CRUSE, A. **Lexical Semantics.** Cambridge: Cambridge University Press. 1986.

ESPINOZA, M.; MONTIEL-PONSODA, E. & PÉREZ, A. P. **Ontology Localization.** 5th International Conference on Knowledge Capture. Redondo Beach, USA, 2009.

FABER, P.; MAIRAL, R. **Constructing a Lexicon of English Verbs.** Berlin: Mouton. 1999.

FABER, P. (ed.). **A Cognitive Linguistics View of Terminology and Specialized Language**. Berlin/New York: Mouton de Gruyter. 2012.

FABER, P.; L'HOMME, M. C. Lexical semantic approaches to terminology: An introduction. Terminology, 20:2. 2014.

FILLMORE, C. J. **Frame Semantics and the Nature of Language**. In: Annals New York Academy of Sciences: Conference on the Origin and Development of Language and Speech 280: https://doi.org/10.1111/j.1749-6632.1976.tb25467.x. 1976.

Fillmore, C. J. **The Case for Case Reopened.** In Syntax and Semantics (Vol. 8). Grammatical Relations. Academic Press, Inc. 1977.

FILLMORE, C. J. **Frame semantics**. Linguistics in the Morning Calm. Hanshin Publishing Co., Seoul, South Korea, 1982.

FILLMORE, C. J. & ATKINS, B. T. S. **Towards a frame-based lexicon**: The semantics of RISK and its neighbors. Frames, Fields and Contrasts: New Essays in Semantics and Lexical Organization, Lawrence Erlbaum Associates, Hillsdale, 1992.

FREIXA, J. **Causes of Denominative Variation in Terminology.** Terminology 12(1): 51–77. 2006.

GAMERO, S. **La traducción de textos técnicos**: descripción y análisis de textos (alemánespañol). Barcelona: Ariel, 2001.

GAUDIN, F. **Socioterminologie: une approche sociolinguistique de la terminologie**. Bruxelles: Duculot. 2003.

GHAZZAWI, N. **Du terme prédicatif au cadre sémantique:** méthodologie de compilation d'une ressource terminologique pour des termes prédicatifs arabes en informatique. PhD thesis presented at the University of Montreal, Montreal. 2016.

GANGEMI, A. **Hybridizing formal and linguistic semantics for the multilingual semantic web.** Proceedings of the 3rd Workshop on the Multilingual Semantic Web. Boston, USA, 2012.

GLENSKI, M. *et al*. **Improving Synonym Recommendation Using Sentence Context**. In: Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, pages 74–78, Virtual. Association for Computational Linguistics. 2021.

GRANGER, S. **The Corpus Approach**: a common way forward for contrastive linguistics and translation studies. *In:* GRANGER, S.; LEROT, J. & PETCH-TYSON, S. Corpus-based Approaches to Contrastive Linguistics and Translation Studies, Amsterdam and Atlanta: Rodopi, 2003.

HARVILL, J.; GIRJU, R., HASEGAWA-JOHNSON, M. **Syn2Vec: Synset Colexification Graphs for Lexical Semantic Similarity**. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5259–5270, Seattle, United States. Association for Computational Linguistics. 2022.

KAY, M. Introduction. *In:* MITKOV, R. **The Oxford Handbook of Computational Linguistics.** New York: Oxford University Press, 2004. p. xvii - xx.

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

LEACOCK, C.; CHODOROW, M. Combining Local Context and WordNet Similarity for Word Sense Identification. In: FELLBAUM, C. **WordNet**: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

LEÓN-ARAÚZ, P.; REIMERINK, A. & ARAGÓN, A. **Dynamism and context in specialized knowledge.** Terminology, 2013.

LEÓN-ARAÚZ, P. & FABER, P. **Context and Terminology in the Multilingual Semantic Web**. *In:* BUITELAR, P. & CIMIANO, P. Towards the Multilingual Semantic Web. Springer, 2014.

LERAT, P. **Vocabulaire juridique et schémas d'arguments juridiques**. Meta 47(2): 155–162. 2002a.

LEWANDOWSKA-TOMASZCZYK, B. **Explicit and tacit**: an interplay of the quantitative and qualitative approaches to translation. *In:* OAKES, M. P. & JI M. Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research. Amsterdam and Philadelphia: John Benjamins, 2012.

L'HOMME, M. C. **Why Lexical Semantics is Important for E-Lexicography and Why it is Equally Important to Hide its Formal Representations from Users of Dictionaries.** International Journal of Lexicography, v. 27. 2014.

L'HOMME, M. C. **Maintaining the Balance between Knowledge and the Lexicon in Terminology: A Methodology based on Frame Semantics**. In: Medical Lexicography and Terminology. Special issue of Lexicography, ed. by P. Peters, J. G. Yongwei, and J. Ding, 4(1): 3–21. 2018.

L'HOMME, M. C. **Lexical Semantics for Terminology**: an introduction. John Benjamins Publishing Company. 2020.

LI, L. **Corpus**. *In:* SIN-WAI, C. The Routledge Encyclopedia of Translation Technology. Routledge: New York, 2015.

MARTINS, M. L. **Equivalências no Dicionário Olímpico**: um fenômeno complexo. Trabalho de Conclusão de Curso (Letras - Inglês) - UNISINOS - Unisinos, São Leopoldo, 2019.

MCCARTHY, D.; NAVIGLI, R. **SemEval-2007 Task 10: English Lexical Substitution Task.** In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 48–53, Prague, Czech Republic. Association for Computational Linguistics. 2007.

MCLUHAN, M. **A galáxia de Gutenberg**: a formação do homem tipográfico; São Paulo, Editora Nacional, Editora da USP. 1962.

MCLUHAN, M. **Os meios de comunicação como extensão do homem**. São Paulo: Cultrix, 1964.

MEL'ČUK, I., *et al*. **Introduction à la lexicologie explicative et combinatoire**. Duculot: Louvain-la-Neuve. 1995.

MICHALOPOULOS, G. *et al*. **LexSubCon: Integrating Knowledge from Lexical Resources into Contextual Embeddings for Lexical Substitution**. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1226–1236, Dublin, Ireland. Association for Computational Linguistics. 2022.

MILLER, G. A. Nouns in WordNet. In: FELLBAUM, C. **WordNet**: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

MITKOV, R. Preface. *In:* MITKOV, R. **The Oxford Handbook of Computational Linguistics.** New York: Oxford University Press, 2004. p. ix - x.

MONTIEL-PONSODA, E.; GRACIA, J.; AGUADO DE CEA, G. & PÉREZ, A. O. **Representing translations on the semantic web**. Proceedings of the 2nd International Workshop on the Multilingual Semantic Web. Bonn, Germany, 2011.

MURPHY, M. L. **Semantic Relations in the Lexicon**: Antonymy, Synonymy and other Paradigms. Cambridge: Cambridge University Press. 2003.

MURPHY, M. L. **Lexical Meaning.** Cambridge: Cambridge University Press. 2010.

NIRENBERG, S.; RASKIN, V. **Ontological Semantics: Language, Speech, and Communication**. Cambrigde: MIT Press, 2004.

PETRUCK, M. R. L. **Frame Semantics**. Handbook of Pragmatics. John Benjamins, 1996.

PIMENTEL, J. **Methodological Bases for Assigning Terminological Equivalents. A Contribution.** Terminology 19(2): https://doi.org/10.1075/term.19.2.04pim. 2013.
PUSTEJOVSKY, J. **The Generative Lexicon.** Language: Linguistic Society of America. 1991.

ROGERS, M. **Multidimensionality in concepts systems**: a bilingual textual perspective. Terminology, 2004.

SAINT-DIZIER, P. & VIEGAS, E. **Computational Lexical Semantics**. Cambrigde: Cambridge University Press, 1995.

SARDINHA, T. B. **Lingüística de Corpus**: histórico e problemática. DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada [online]. 2000, v. 16, n. 2.

SINCLAIR, J. **Trust the Text**: Language Corpus and Discourse. London: Routledge, 2004.

TALMY, L. **Toward a Cognitive Semantics.** Cambridge: MIT Press. Language: The Journal of the Linguistic Society of America. 2000.

TUGGY, D. **Ambiguity, polysemy, and vagueness**. *In:* DIVIJAK, D. Cognitive Linguistics. De Gruyter Mouton. 1993.

YEE, O. K. O. **Natural Language Processing**. *In:* SIN-WAI, C. The Routledge Encyclopedia of Translation Technology. Routledge: New York, 2015.