Universidade de Lisboa
Faculdade de Letras

Multiword Proper Nouns in Multilingual Glossaries for Machine Translation and Post-Editing

Mestrado em Tradução

Nadezhda Andrianova Metodieva

Relatório de estágio especialmente elaborado para a obtenção do grau de Mestre, orientado pela Professora Doutora Sara Gonçalves Pedro Parente Mendes e coorientado pela Professora Doutora Helena Gorete Silva Moniz

2023

Acknowledgements

I would like to express my sincere gratitude to the following people whose support and guidance have played a crucial role in my academic journey:

First and foremost, I express my deepest love to my amazing mother Mary and my dear grandparents Temenuzhka and Simeon. Their constant support and boundless love have been the pillars of strength, courage and wisdom throughout my life. In addition, I am grateful to Elena, Stefan, Ventsislava, Adriana and Monica for all their love and support.

I am deeply grateful to Professor Sara Mendes, who believed in me all along. Her extraordinary patience, invaluable suggestions and guidance have not only shaped the course of my work, but have also been a constant source of inspiration. Her support, both academic and emotional, over the years is immeasurable. I am equally grateful for the moments of laughter we shared at our meetings as we talked about the imaginary characters and delved into the captivating world of proper nouns in video games and literature. It is impossible to adequately convey the depth of my gratitude for her mentorship.

I am indebted to Professor Helena Moniz for her insightful suggestions and for making my internship possible. Her efforts in organising the internship, addressing authentic translation issues in a real-life setting and fostering invaluable partnerships have greatly enriched my academic and professional journey.

I would like to thank Rui Santos for encouraging and supporting me from the beginning and for showing me the light so often, even when I forgot to look for it.

I am grateful to all the remarkable people at Unbabel who have supported me in various ways. My special thanks go to Catarina Silva, Marianna Buchicchio, Nikita Savytskyi, Carla Parra Escartín, Vera Almeida and Camila Pohlmann.

My heartfelt thanks go to Tatiane Rocha for her tireless support and the friendship that developed during this internship.

I am truly grateful to Vânia Marlene Marinho and her family, who have always stood by me and opened their hearts to me since my arrival in Portugal.

I sincerely thank Susana Correia and Gil Duarte for the countless inspiring conversations and wonderful afternoons we spent together and for motivating me in so many difficult moments.

I extend my gratitude to my friend Joana Rebelo for enriching my life immeasurably with her support and companionship.

My special thanks to Yasmina Habib, whose understanding, patience and kindness were a source of comfort in the final stages of completing this thesis.

Last but not least, I would like to thank from the bottom of my heart all my wonderful friends and family in Bulgaria and Portugal, especially Ana-Maria Mladenova, Didi Karaleeva, Bozhidara Delyakova, Antoniya Serafimova and Elvira Murati - for their enduring friendship and support throughout the years. Their presence in my life has been a source of joy and inspiration.

Thank you all for your unwavering support and belief in me throughout this journey. Your presence has made this achievement possible.

Nana Metodieva

2023

Resumo

Este relatório baseia-se no trabalho desenvolvido durante um estágio na Unbabel, uma empresa que utiliza tradução automática e pós-edição de textos traduzidos automaticamente. O estágio envolveu tarefas e responsabilidades nessa área e proporcionou uma experiência envolvente de colaboração. Durante o estágio, foi possível compreender em profundidade as ferramentas e recursos utilizados nas tarefas de pós-edição de tradução automática, bem como fazer sugestões para aprimorá-los. Essa experiência despertou o meu interesse em aprender mais sobre os desafios da tradução de glossários multilingues, com foco especial na tradução de nomes de personagens multipalavra em videojogos.

Os glossários multilingues desempenham um papel fundamental no fluxo de trabalho da Unbabel, facilitando a pós-edição de textos traduzidos automaticamente e garantindo a coerência terminológica nos projetos de tradução para clientes específicos. Os processos de criação e tradução dos glossários são realizados manualmente, sendo crucial compreender as prioridades e desafios envolvidos, e a maneira como o sistema de tradução automática e os pós-editores os utilizam. Um glossário bem elaborado simplifica o trabalho dos pós-editores e assegura traduções de melhor qualidade.

Este relatório está dividido em seis capítulos. O Capítulo 1 é uma introdução que fornece informações sobre a empresa onde decorreu o estágio e a organização do texto. O Capítulo 2 contém um relatório detalhado sobre as atividades realizadas durante o estágio, incluindo uma descrição das tarefas executadas, os processos de trabalho na empresa, a funcionalidade das ferramentas de tradução e de apoio aos tradutores e os recursos disponíveis para a equipa de tradução. O estágio abrangeu testes de ferramentas para pós-edição de tradução automática, anotação de erros linguísticos em textos de chegada e avaliação de traduções. Também incluiu a utilização e análise de recursos linguísticos, tais como glossários multilingues e guias linguísticos. Isto proporcionou uma compreensão aprofundada de todos os aspetos relacionados com os glossários na Unbabel.

Algumas conclusões importantes emergiram da experiência de estágio. Em primeiro lugar, proporcionou uma compreensão aprofundada das ferramentas de controlo de qualidade e de pós-edição, bem como dos seus propósitos e funcionalidades. Em segundo lugar, o ambi-

ente colaborativo incentivou a troca de ideias e o desenvolvimento de pensamento crítico. Mais relevante ainda no âmbito deste relatório de estágio foi a tarefa de curadoria do glossário que proporcionou considerações valiosas sobre aspetos importantes da criação e tradução de glossários. Esta parte inclui reflexões sobre a estrutura das unidades do glossário, identificação de padrões nas unidades de partida e observações sobre como esses padrões foram tratados nas unidades de chegada. Adicionalmente, foram feitas algumas considerações sobre a informação incluída no glossário e a necessidade de informação adicional.

O Capítulo 3 estabelece um enquadramento teórico relevante para este estudo. Começa por examinar a evolução da tradução automática e das ferramentas de tradução assistida por computador, fornecendo uma análise das abordagens de gestão de terminologia e glossários. Além disso, reflete sobre expressões multipalavra para fins específicos e com particular ênfase nas unidades de glossário que englobam nomes próprios multipalavra. Destaca-se igualmente a distinção entre glossários concebidos para uso humano e aqueles destinados a serem utilizados por sistemas de tradução automática. Este enquadramento teórico fornece base necessária para a análise dos aspetos práticos deste estudo.

O Capítulo 4 descreve a metodologia utilizada para a curadoria do glossário, bem como os critérios de seleção das unidades do glossário para análise. O propósito da curadoria é analisar a estrutura do glossário e das suas unidades, identificando desafios de tradução. Este capítulo também aborda questões relacionadas com a anonimização de dados, em conformidade com a regulamentação de proteção de dados. O processo inclui a recolha de dados, a descrição de alguns dos recursos utilizados, a seleção das unidades do glossário a serem analisadas e os critérios para tal seleção, além de uma análise preliminar que abrange vários aspetos, como a estrutura das unidades e algumas considerações ortográficas. A seleção das unidades do glossário prevê uma análise de unidades que podem apresentar desafios significativos para a tradução automática e a pós-edição humana, nomeadamente unidades compostas por várias palavras. Foi identificado um nível adicional de complexidade em relação a um grupo específico de unidades compostas por várias palavras que corresponde a unidades que se referem a nomes de personagens presentes nos videojogos do cliente pelo que estas foram escolhidas para análise adicional.

O Capítulo 5 é o núcleo deste relatório de estágio e concentra-se na análise das unidades de glossário escolhidas. Estas unidades incluem um conjunto de 28 unidades de partida em inglês e as suas respetivas traduções para para o búlgaro, o português europeu e o português do Brasil. A análise abrange várias dimensões, incluindo a observação de sequências lineares de categorias morfo-sintácticas dos elementos constituintes das unidades de glossário selecionadas. Adicionalmente são consideradas questões relacionadas com o género, a ordem das palavras e

a definitude. Neste capítulo, também são abordados os aspetos criativos observados nos nomes próprios multipalavra, incluindo os efeitos fonéticos que podem influenciar o processo de tradução.

A nossa análise revela que a estrutura morfo-sintática das unidades de partida analisadas segue consistentemente o padrão [NOME PRÓPRIO] + [DESCRIÇÃO DEFINIDA]. Em contraste, a estrutura das unidades de chegada em búlgaro e português mostra variações consideráveis. Procede-se à comparação das estruturas das unidades de partida e chegada e examinam-se as estratégias de tradução adotadas. Adicionalmente, são consideradas as mudanças semânticas resultantes das alterações na ordem das palavras.

Além disso, no Capítulo 5, explora-se a presença de efeitos fonéticos nas unidades do glossário e a sua importância, especialmente no contexto de nomes de personagens de videojogos. Os efeitos fonéticos, como aliteração, rima e aliteração onomatopaica estão presentes em mais de metade das unidades analisadas (57%). Estes efeitos fonéticos não se limitam apenas à literatura, mas ocorrem também em lemas, jogos de palavras, rimas infantis e letras de canções. No nosso caso, ocorrem em nomes de personagens de videojogos e têm como objetivo aumentar a memorização dos nomes das personagens, tornar os jogos mais envolventes e estabelecer uma ligação entre as personagens ao universo único de cada jogo. Os jogos considerados neste trabalho apelam a um público que aprecia os aspetos lúdicos e criativos de vários elementos do jogo, incluindo os nomes das personagens. Consequentemente, os tradutores devem descobrir todos os significados e efeitos incorporados nas unidades de partida e refleti-los nas unidades de chegada. No entanto, muitas vezes não é possível transferir todas as nuances e facetas da unidade de partida para a unidade de chegada.

Este capítulo destaca a relevância do contexto no processo de tradução, em particular ao atribuir género a nomes próprios multipalavra e a considerável quantidade de tempo e pesquisa que muitas vezes é necessária para fundamentar decisões de tradução nesse contexto. Atribuir género a nomes próprios é particularmente desafiante para os tradutores, devido à natureza das línguas de partida e de chegada. Os tradutores têm de atribuir género de forma deliberada às unidades de chegada ou reconciliar diferenças de género não intencionais (ou seja, o género inerente dos substantivos em búlgaro e português com o género percebido de nomes próprios convencionais). Neste processo, frequentemente é necessário investigar fontes públicas sobre os jogos para confirmar ou determinar o género dos nomes próprios pouco comuns. Isso sublinha a importância do contexto no processo de tradução, bem como a quantidade significativa de tempo e esforço que os tradutores devem alocar à pesquisa, quando os criadores de glossários não fornecem contexto suficiente. Além disso, ressalta os potenciais erros que podem ocorrer se os tradutores não tiverem tempo ou acesso para pesquisar a informação indispensável.

O Capítulo 6 é reservado às considerações finais do relatório de estágio e inclui algumas recomendações para os criadores e os tradutores de unidades de glossários que incluem nomes próprios multipalavra em glossários para tradução automática e pós-edição de outputs de tradução automática.

Em resumo, este relatório de estágio proporciona uma análise aprofundada da curadoria de glossários multilingues e dos desafios associados à tradução. Também aborda questões sobre o universo dos nomes multipalavra de personagens de videojogos. As perceções, recomendações e possíveis direções para investigações futuras aqui apresentadas contribuem para o diálogo contínuo sobre a gestão de glossários multilingues e têm o potencial de melhorar a qualidade da tradução em áreas especializadas, tais como a tradução de conteúdo relacionado com videojogos e a sua localização.

Abstract

This report is based on work carried out as part of an internship at Unbabel. As Unbabel is a company that uses machine translation and post-editing of machine translation outputs, the internship involved tasks and responsibilities in this area. This included testing of tools for post-editing of machine translation outputs, annotation of linguistic errors in target texts and evaluation of translations. It also included the use and analysis of linguistic resources such as multilingual glossaries and language guidelines. This report is divided into six chapters. The first chapter is an introduction containing information about the company where the internship took place and the organisation of the work carried out. The second chapter contains a detailed report on the activities carried out during the internship, including a description of the tasks performed, the work processes in the company, the functionality of the translation tools and the resources available to the translation team. Chapter 3 presents a theoretical framework relevant to the topics addressed and analysed in this work. The topics selected cover issues related to machine translation and computer-aided translation tools and provide a historical context for these technologies. This chapter also examines the field of terminology management and glossaries, focusing on multiword glossary units, especially multi-word proper nouns. Chapter 4 describes the selection process for a particular group of glossary units chosen for in-depth analysis. The criteria for selecting the glossary units are described and the context in which these units find their place in the analysis is explained. Chapter 5 is devoted to the analysis of the translations of a number of multiword proper nouns in English into Bulgarian, European Portuguese and Brazilian Portuguese. This analysis covers several dimensions, including the observation of linear sequences of morpho-syntactic categories corresponding to the structure of the selected glossary units. Considerations regarding gender, word order and definiteness are also included. In addition, this chapter deals with creative aspects identified in multiword proper nouns, including phonetic effects that may influence translation decisions. Chapter 6 is reserved for the final reflections regarding the work developed and includes some recommendations for translators of multiword proper nouns, which appear in glossaries for machine translation and human post-editing of machine translation outputs.

Palavras-Chave
Keywords

Palavras-Chave

Tradução Automática

Pós-Edição de Tradução Automática

Glossários Multilingues

Tradução de Glossários

Nomes Próprios Multipalavra

Keywords

Machine Translation

Post-Editing of Machine Translation

Multilingual Glossaries

Translation of Glossaries

Multiword Proper Nouns

Contents

List of Figures

List of Tables

List of Abbreviations

ADJ - Adjective

ART - Definite Article

CAT - Computer-Aided Translation

CSS - Customer Support Service

G0 - A set of glossaries of a client to be incorporated in a single glossary

G1 - First version of the analysed glossary (working version)

G2 - Second version of the analysed glossary (final version)

LSP - Language for Special Purposes

MQM - Multidimensional Quality Metrics

MT - Machine Translation

MWE - Multiword Expression

MWEs - Multiword Expressions

MWESP - Multiword Expressions for Special Purposes

NC - Noun Compound

NCs - Noun Compounds

NLP - Natural Language Processing

NMT - Neural Machine Translation

PoS - Part of Speech

PREP - Preposition

P.NOUN - Proper Noun

QT21 - Quality Translation 21

RBMT - Rule-Based Machine Translation

SL - Source Language

SMT - Statistical Machine Translation

ST - Source Text

STs - Source Texts

TL - Target Language

TM - Translation Memory

TMS - Terminology Management System

TMs - Translation Memories

TT - Target Text

TTs - Target Texts

1 Introduction

This dissertation builds on the work developed and data collected under the scope of an internship at Unbabel. It presents the tasks and activities accomplished during the internship, carried out as part of the curriculum of the Master's program in Translation of the Faculty of Arts and Humanities of the University of Lisbon, Portugal. Generally, the aims of a curricular internship are, firstly, for the intern to get familiar with a working environment, to put into practice knowledge obtained during the curricular part of the studies, hence becoming more competitive on the work market, and secondly, to obtain the Master degree in Translation, as stipulated in the rules of the program (Faculdade de Letras da Universidade de Lisboa, 2018), after the elaboration and public discussion of an internship report. In compliance with the rules in force for the academic year of 2017/2018, the internship was divided between the two semesters of the second year of studies and included a total of 240 working hours at a company. In my case, the company where the internship was carried out was *Unbabel Ltd.*

Specifically, the main goals of the internship at Unbabel were: getting to know the company and the translation market; getting used to an international technological company environment; accomplishing tasks that could be useful for the development of the company; learning how to work in a team and how to take part in the company's community and culture; providing important results based on a thorough scientific analysis, concerning the topic of the internship report.

In order to provide quality service, one of the fundamental steps in the Unbabel workflow consists in creating a functional and precise glossary. The aims of the glossaries at the company include aiding the post-editing of Machine Translation (MT) outputs and guaranteeing coherence regarding the terminology used in translation projects accomplished for a certain client. This process is manually developed. Therefore, it is important to understand the priorities in the creation of the glossary, the way it is processed by the MT system and how it is going to be used by the post-editors. A good glossary makes the editors' work easier and guarantees higher quality translations to be provided to the customers.

The choice to dedicate this work to the analysis and improvement of the company's glos-

saries is based not only on their importance in the Unbabel workflow but also on my personal interest in building and translating glossaries, organising terminological databases and their use in professional human translation and in MT post-editing.

The internship experience offered a basis for an in-depth understanding of all aspects of the glossaries at Unbabel. This included a thorough comprehension of the company's post-editing and quality assurance tools and the resources involved in the translation processes. These are described in Chapter 2, along with the activities carried out during the internship. The chapter is divided into three main sections: work arrangements during the internship and brief information about the company; description of the workflow and the glossaries used at the company; description of the completed tasks, which include testing of the tools of the company (such as Post-Editing, Annotation and Evaluation tools), analysis of the linguistic resources, preparation of feedback and suggestions and curation of a glossary of a client.

Chapter 3 establishes a theoretical framework by examining the evolution of machine translation and computer-aided translation tools and providing insights into their operation and use. It also addresses the field of terminology management and the glossaries used by human translators and machine translation engines, with particular attention to multiword glossary units, especially multiword proper nouns. This theoretical background gives us the necessary understanding to analyse and explore the practical aspects of our study.

Chapter 4 focuses on the process of selecting glossary units for further analysis and describes the criteria used. This chapter presents the process of curating the glossary, some of the resources used and some aspects of data anonymisation.

Chapter 5 is the main part of our research. It is dedicated to the analysis of the source and target glossary units, which are composed of multiword proper nouns of video game characters. These proper nouns are translated from English into Bulgarian, European Portuguese and Brazilian Portuguese. The analysis covers several dimensions and includes an observation of the linear sequences of morpho-syntactic categories within the structures of the glossary units. The chapter is divided into four main parts. The first part addresses the analysis and comparison of the morpho-syntactic structures of the units and includes considerations of their definiteness and word order. The second part is devoted to some findings related to orthography and punctuation. Considerations on the gender of the selected glossary units are included in the third part. The fourth part of Chapter 5 describes some phonetic effects that occur in the units, their importance and influence on the translation process, and provides additional context about the type of video games in which the characters' names appear.

Finally, Chapter 6 addresses the concluding remarks on the internship report. It summarises the findings and provides recommendations for glossary creators and translators dealing with multiword proper nouns in glossaries for machine translation and post-editing of machine translation outputs.

2 Activities Report

In this chapter, there is a short description of the company which hosted my internship, in Section 2.2, its translation workflow and of different tools linked to the quality assurance of the translations produced; a description of the multilingual glossaries, used at Unbabel, their structure and functionality is provided in Section 2.2.2; in Section 2.3 the responsibilities and the tasks completed during the internship are described and some conclusions are presented in Section 2.4.

2.1 Work arrangements

The internship at Unbabel Ltd. started on September 19, 2017 when the first meeting between the interns, their supervisor in the company and their academic advisors took place in the company's office in Lisbon, located at *Rua Visconde de Santarém*, 67B. A general overview of the workflow at Unbabel, as well as provisional thesis topics were presented by Helena Moniz, PhD, a Quality Assurance Researcher at Unbabel and supervisor of the four interns hosted by the Quality team. The interns referred to in this text are: myself - Nadezhda Metodieva, Tatiane Oliveira, Catarina Xavier and Rhandra Lopes, all studying in their second year of the Master's in Translation at the Faculty of Arts and Humanities of the University of Lisbon at the time of the internship.

It was decided that the interns would come to the office twice a week, for a total of eight working hours a week. This allowed them to attend the company's two weekly events to be integrated into the company culture and learn about the processes, work routines and newly developed tools and software, as well as some statistics about the performance and goals of the different teams. The events mentioned included: a weekly lunch followed by a presentation of the company's progress and a weekly general meeting with information about the teams' main goals for the following weeks or months.

The internship can be roughly divided into two parts: a first part concentrated on teamwork and a second part focused on the individual topic assigned to each intern. The specific tasks

performed during these periods are presented in section 2.3.

2.2 Unbabel – translation workflow and glossaries

Unbabel Inc. is a translation company that operates with human-edited MT. Founded in 2013, it is headquartered in San Francisco (USA) with hubs in Lisbon (Portugal), London and Edinburgh (UK), New York and San Francisco (USA), Timișoara (Romania) and Cebu (Philippines). Its main human resources are distributed between the various offices and a big freelance community, working remotely.

Since the early years of research in MT in the 1950s, MT has never reached the same quality as a translation made by a professional human translator. According to Hutchins (1986) the MT goal of the 1960s, obtaining a fully automatic high quality translation, has changed over time and many researchers in the field consider that "[if] a better quality is required then collaboration of man and machine is essential".

According to its official web page at the time of the internship, Unbabel uses Neural Machine Translation (NMT) refined by a community of 50,000 bilinguals in 28 languages that edit the output of the MT system (Unbabel, 2019).

As professional translations are costly and time-consuming, most multinational companies prefer to have only part of their content translated by professionals. Nevertheless, there is a large amount of text that never reaches professional translation agencies, such as customer service emails, chat interactions or internal company communications. This content is usually translated using MT, despite the often unsatisfactory quality of the output text. In addition, some texts need to be translated in near real-time (e.g. chat interactions between customers and customer service representatives), very quickly (customer service emails) or without spending much time on them because only a general idea of the text is needed.

The translation solutions offered by Unbabel are used in the cases mentioned above.[1] However, for more specialised texts, such as medical, legal or technical texts, where thorough accuracy is required, and for literary works that need to be translated with creativity and sensitivity to cultural differences, it is still obligatory to use professional human translations.

To better understand why the company focuses on translating the aforementioned type

[1] This and all other statements about the company, the tools used and the workflow are based on the information available during the internship. Since then, the company has expanded its services, but we have not received any information about the nature of these new services, the tools developed or the workflow.

of text, first, it is important to have in mind the profile of the post-editors at Unbabel, who are not professional translators in most cases, but bilinguals that are proficient in the source language (SL) and are usually native in the target language (TL). In general, the professional translators of legal, medical or technical texts are fluent or native in the SL and the TL and are also well familiarised with the special field and the terminology that is used in their assignments. The post-editors at Unbabel are not required to have background in any special field, nor to prove their proficiency in translation by having some experience or academic background in this area.

Second, the question of the authorship of the translation is also an aspect that must be taken into account. In the case of legal and medical texts, it is generally the translator who proves his or her identity, certifies the accuracy of the translation and assumes responsibility in the event of discrepancies. With Unbabel, the texts are first machine-translated and then divided into segments and distributed to different post-editors whose identity remains anonymous. Therefore, responsibility for the translation produced cannot be attributed individually to any of the participants in the process.

Third, the MT system of Unbabel has not been trained to translate technical texts, although it could be if it is given corpora large enough for training.

These considerations focus on the accuracy and responsibility of the translation, as well as the possibility of errors. To address these points thoroughly, we need to take a close look at the translation workflow.

2.2.1 Translation workflow

The internship provided an opportunity to understand first-hand the translation workflow and the involvement of people and machines in the complex translation process. To understand this and lay the groundwork for the main topic of this thesis, let us look at the company's translation workflow.

Before Unbabel starts delivering translations to a client, the company applies a specific model for on-boarding new clients, which includes the following steps: understanding the client's vision of how they want to present themselves to their customers and defining the register (formal/informal) to be used in the translations; manually creating a multilingual glossary using the client's specialised terminology; gathering more information about the client's business to select the best editors from the Unbabel community for the tasks at hand.

At this point, the client's glossary is created at Unbabel using the linguistic resources submitted by the client and any public information available for that client. In this way, candidate terms are identified and selected for the glossary. Then the source terms are translated into the languages the client works with or intends to work with.

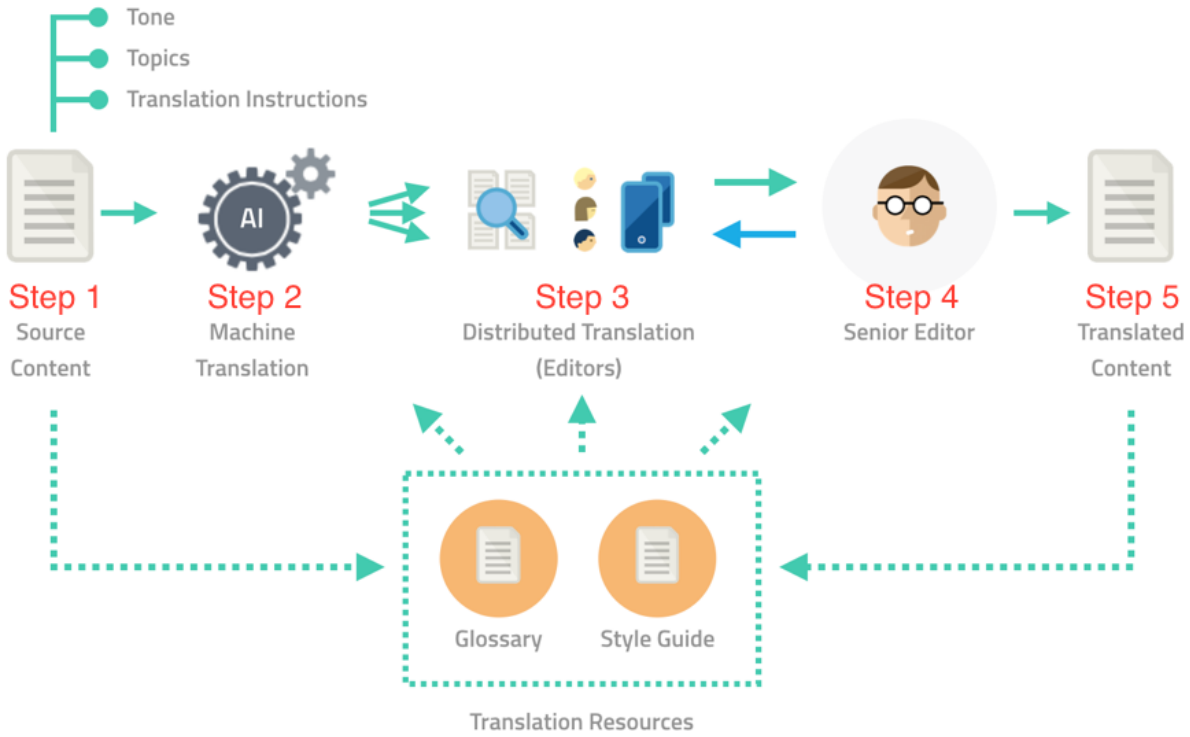

Figure 2.1: A model of the Unbabel translation workflow.

Figure 2.1[2] shows a model of the Unbabel translation workflow, which is described in more detail below (Unbabel, 2018c). It is important to mention that some of the steps described are not automatic and involve human intervention.

After the on-boarding process, the client sends the source texts to Unbabel ("Step 1"). These are firstly processed by a system that detects and deletes all the extra white spaces (this is important for the correct application of translation memories and glossaries) and anonymises all sensitive and personal data.

Next, a dedicated glossary system automatically searches for equivalences between the words in the source text (ST) and the terms in a previously created glossary. If it finds any, it directly replaces a source term with a target term contained in the glossary and highlights both in the platform for post-editing. Typically, glossaries are created in English and then translated into

[2]Adapted from the information available at the Unbabel website: https://developers.unbabel.com/docs/getting-integrated

multiple languages. This poses particular challenges, which we will analyse below (see section 3.3.2).

Afterwards, another tool looks for exact matches in Translation Memories (TMs) that have been built using previous translations made at Unbabel for the same client, if any. The system that processes TMs looks exclusively for exact character match, which means that it would not detect a sentence that has the same meaning as the sentence being translated, but a different word order, for example, or if synonyms are used.

TMs consist of sentences and expressions that have already been verified by a human editor, therefore their reliability is high. If an exact match is found, the segment is applied directly into the target text (TT). The quality of the Translation Memory (TM) segment is not re-evaluated at this point.

If the glossary terms are not correctly identified in the TMs (e.g. due to differences in the inflected forms of the source term) or are not replaced by their target equivalents, and/or do not appear in the TM segments due to technical problems, these segments are discarded. This shows that the glossaries take precedence over the TMs.

Once TMs and glossaries have been applied to the text, it goes through the MT system ("Step 2"), which translates the content for which no exact match was found in the TMs or glossaries. The MT system is "blind" for the glossary terms, i.e. it does not take their existence into account and translates the parts of the text before and after each term. Therefore, it is important to emphasise that one of the characteristics of an efficient glossary is that it is concise, as glossaries that are too extensive reduce the probability of successful MT. Like any state of the art NMT system, the MT system used at Unbabel considers a source sentence as a whole in order to produce a target sentence. If a part of the text is missing, e.g. a glossary word, the structure of the phrase is changed, altering the linguistic context of the words, which degrades the performance of the MT system.

Subsequently, after target terms found in the glossary have been inserted into a MT output, another software examines the different output variants offered by the MT system. It chooses the best one by using different grammatical rules (e.g. agreement) and the context within the phrase. For example, suppose that the following phrase is used as a source: "The lipstick will be sent to your home within three days" and the MT system suggests the following two variants for translation into Portuguese: 1. *O batom será enviado para a sua casa dentro de três dias* and 2. *A batom será enviado para a sua casa dentro de três dias.* This software would prefer variant 1 to the ungrammatical number 2, in which the feminine form of the article is used.

After that, another system makes a quality estimation. This is a value that is automatically assigned to a MT output in terms of its estimated quality. If an output ensures a reliability rate above 90 % (using predefined rules that were not communicated to the interns), the system automatically sends it to the client. If this threshold is not reached, the MT is sent to the human post-editors for improvement ("Step 3").

In Step 3, the text is divided into segments, i.e. small fragments of text, which are distributed in parallel to different post-editors of the respective language pair. This ensures a quick delivery of the translation, as each editor can review a small part of the text. They can even do this on their mobile phones, wherever they are (the company provides a smartphone application so that post editors can work on their mobile phones).

Editors have various tools available to support the post-editing process. On the editors' platform, they can see the source text, the machine-translated target text that can be edited, the client's instructions regarding the register (formal/informal), the requirements to be taken into account for the translation (e.g. whether it should be approximately the same length as the ST) and the domain of business of the client. It also highlights the parts of the text that have been taken from TMs and the glossary terms. There is another useful tool called *Smartcheck* that alerts the editor when it detects possible problems in the text (e.g. non-compliance with glossary terms, typos, agreement issues, etc.).

The edited text segments are then combined and another editor, called *Senior Editor* – an editor with good ratings and more experience on the platform, reviews the full text resulting from the combination of the segments, looking for inconsistencies regarding lexical selection, punctuation, etc. and corrects wrongly translated words or phrases ("Step 4"). After checking and improving the quality of the text, the Senior Editor sends it directly to the client via special tools and applications integrated into the client's software ("Step 5").

Unbabel also has a community of professional linguists, called *Evaluators*, who assess the performance of editors and help select the best editors to make translations for the company. There is also another team of linguists, called *Annotators*, who annotate the translations with qualitative and quantitative information regarding the errors to help improve the automatic systems and the workflow.

2.2.2 Glossaries at Unbabel

As explained in the previous section, glossaries are tailored to the needs of each client and are manually created by members of the Unbabel team or the community of linguists. Unbabel has

three ways of determining which terms to include in the glossary. The first is to collect candidate terms provided by the client from the internal glossaries. The second one is to manually select what to include based on the client's website or other content available online or submitted by the client. The third way involves an automatic scraping of the most frequently used words in the texts provided by the client for translation. The latter strategy allows for the inclusion of terms that were missing in the client's internal glossary (e.g. if the terms only began to be used after the client's own glossary was completed). However, this is a rather inefficient strategy in most cases, as it favours quantitative frequency, which is higher for general vocabulary. General vocabulary is not specific and is covered by MT. The inclusion of lexical units that are not essential may contribute to the creation of less concise glossaries, which may have the effect already described on page 9. All three methods can be used to create a single glossary that covers all the terms that need to be included.

2.2.2.1 Purpose and main priorities of the glossaries

Glossaries ensure consistency in the translations throughout different segments of the same text and in different assignments for the same client. The terms included in the glossary are automatically recognised in the input text and their equivalents (according to the glossary) are inserted in the output. They are also highlighted in the Editor's platform during the post-editing process, as well as in the platforms available for Evaluators and Annotators.

Glossaries not only contribute to consistency, but also aim to make the translation process time-efficient, thereby reducing the cost of translations, as the various human elements in the pipeline have to spend less time researching a particular term in the ST, for example. For example, if there is a term in the ST that is already in the glossary, the next person in the pipeline (the Editors, the Senior Editor, the Evaluator and/or the Annotator) simply has to accept the word as correct without having to spend further time researching it. Of course, this is only the case if the terms in the glossary are chosen wisely. A small amount of additional time still has to be spent in case the inflected forms of the target term need to be changed. This is done manually each time a glossary term appears in the TT.

In this context, it is important to emphasise that Unbabel's glossaries are used by both humans and machines. Therefore, their characteristics reflect the needs of human post-editors, but also take into account the fact that they are processed by a MT engine and other Natural Language Processing (NLP) software. On the one hand, it is useful that the glossary units are automatically retrieved and appear directly on the TT's screen, so that they can be edited. On

the other hand, there are limitations and some terms that are part of the glossary can be missed by the system if they have different inflected forms than those included in the glossary. In this case, the glossary units are not retrieved by the system and do not appear on the screen of the post-editor, who has no way of accessing them and using the information in the glossary.

According to Melby (as cited in Cabré and Sager, 1999, p.168):

> Human-oriented and machine-oriented terminology files contain quite different information and a machine-translation dictionary will contain common words not needed in a human-oriented terminology file, but ideally, they will be synchronised to contain records for the same technical terms. This will allow the human-oriented files to contain the documentation of how the translation was chosen and how the term is used by human experts, while the machine-oriented files contain the syntactic and semantic codes needed for a machine processing.

In the case of the glossaries used at Unbabel, we can say that they are more machine-oriented and less human-oriented because of the way they are accessed. The emphasis is on the fact that they are applied before MT and offer the possibility of subsequent editing by a human translator. However, since the post-editors do not have access to the full glossary, they cannot search for terms or parts of terms that occur in the ST with different inflected forms, nor can they perform a reverse search, i.e. search the database for the target term. It is also not possible for them to search based on the description, part of the description or the domain. Possibilities, that are available in some glossaries created exclusively for human use.

Unlike glossaries made specifically for humans, the ones used both by machines and humans must be very concise. In other words, they must contain only the specific terminology of the client, without ambiguous words. Thus, it may be vocabulary from the customer's business area (e.g. financial services: *transaction*, *purchase order*); specific vocabulary whose translation has already been selected by the client (e.g. *e-mail address* to be translated as *endereço eletrónico* in Portuguese in all the client's texts); vocabulary that must be retained in its original form (e.g. copyrighted brand names); vocabulary that needs to be transliterated or phonetically transcribed in a specific way in languages that use a different alphabet (e.g. proper nouns), vocabulary that contains abbreviations, and so on. All these cases consist predominantly of multiword units in the glossaries at Unbabel.

As mentioned above, general vocabulary is sometimes included in glossaries; however, this should be avoided for several reasons. First, having in mind that the MT system considers glossary terms as blank spaces (slots) during the translation process, the more blank spaces there are in an input sentence, the worse the MT system is going to perform because it is trained to

translate sentences as a unit, not fragments of sentences. Moreover, the MT engine translates general vocabulary considerably well and inflects the lexical units that have distinct morphological features (which would not be the case if they were included in the glossary). Second, general vocabulary should remain outside the glossary in order to avoid incorrect translations. For instance, some of the words selected for a glossary are polysemous or can have homographs. However, it is hardly ever the case that the translation of a pair of homographs in one language consists of a pair of homographs in another. That is why it is logical to include all these homographs and to have some kind of mechanism to distinguish which translation is the correct one with regard to the specific context of the ST (the same way some MT systems use the context of occurrence to distinguish between two homographs). Yet, the glossaries at Unbabel must include just one target term for each source term, and there is no software that could select the correct translation if there were more than one target term per source term.

2.2.2.2 Structure and functionality

Unbabel has a special platform that allows the team responsible for managing and developing glossaries to create them online, link them to the Unbabel translation pipeline and view all available translations. In this way, different translations can be added to the glossaries, updated as needed, linked to the client's account and triggered when the client sends new translations.

However, for the purposes of this study and during the internship, I worked with offline glossaries in Excel format, containing the same information that was available on the platform at the time. This was the case because the task I was given involved curating the glossaries and I did not use them as a post-editor on the platform.

A description of the template used by the company as well as a detailed analysis of these Excel spreadsheets is provided in Chapter 5.

The way the glossary system worked at the moment the internship took place is simple and very straightforward, but it met the needs of the company. As mentioned earlier, the glossary system searches its database for matching terms in the ST provided for translation, and it only recognises exact matches. This is quite important because it cannot recognize a part of a word, such as a lemma with different inflexions. Also, multiword glossary terms are searchable only as a whole, which means that if one of their components can also be used as an isolated glossary term, it must be added in a separate entry so that the system can take it into account. Furthermore, in this case, the system first searches for a match of the longest character combination, while the shorter ones are only taken into account if no match is found. For example, if a glossary contains

terms such as *Customer Service Centre*, *Customer Service* and *Customer*, the system first looks for the longest combination, i.e. *Customer Service Centre*.

The principle of exact match also applies when it comes to punctuation in the glossary. For example, if words with punctuation marks have been extracted during the automatic extraction of terms, the system considers the punctuation marks as part of the word. If this word occurs with other punctuation marks or without any, the system does not recognise it in the ST and therefore does not treat it as a glossary unit. This is a clear example of the importance of curating glossaries and removing such entries from this type of resource at Unbabel. Apart from this, there are some rare cases where punctuation marks can serve as disambiguation factors, e.g. when a common word is used as a specific vocabulary of the client only if there is an exclamation mark next to it.

The capitalisation of glossary terms is also important. The system that processes glossaries and identifies terms in a ST is able to recognise them even with a different capitalisation. Nevertheless, it offers several options for capitalisation that highlight how important encoding decisions are in this aspect. For instance, there is the possibility of having the system capitalise a particular glossary unit in the same way as its equivalent is capitalised in the ST. Another option is to keep the capitalisation of the target term as it is indicated in the glossary and to disregard the way it is capitalised in the ST. These two options are available when the glossary is created and translators can choose the best option for each glossary term. This method is useful because, for example, in German all nouns should start with a capital letter, which is not the case in English. This way, in the glossaries, the nouns in the German target units begin with a capital letter, these are then applied directly into the TT, which saves time in post-editing. In addition, there are cases where some companies, for marketing reasons, prefer their products to always be capitalised in a certain way. Thus, the desired capitalisation can be defined in the glossary so that it is applied every time the name of the product is used.

According to Unbabel's guidelines for creation and translation of glossaries,[3] ambiguity should be avoided. This means that homographs should not be included in glossaries (as mentioned on p. 13) unless one of the forms is a very specific expression used by the client in a very consistent way. For example, if we consider the sentences *He is going to test the software* and *The English test was way too easy for her*, in which the word *test* is a verb and a noun respectively, this word should not be included in the glossary. As mentioned earlier, the reason for this rule is simple: there cannot be two glossary entries with the same source term because the system

[3]Internal report, not publicly available.

cannot decide which one is best in the specific context of the ST. For the same reason, there can be only one translation per language for each source term. Bowker and Ciro (2019:46) define homonyms as "linguistic accidents" and explain the difficulties of MT engines in dealing with them and with other types of ambiguity, such as "structural ambiguity".

Apart from the problems associated with ambiguity, it is important to note that glossaries are used in both ways depending on the SL of the input. For example, there could be a glossary that contains terms in English and Portuguese. The English terms are used as source terms when translating from English to Portuguese and as target terms when translating from Portuguese to English (this topic is discussed in more detail in section 3.3.2). In this case, if the target terms are used as source terms, the same rule should apply - there should be no repeated target terms, as the glossary system cannot decide which one to use.

At the moment of the internship, glossaries were tailored to the customer. However, one can assume that in the future there may be glossaries per domain. For example, a general glossary containing all terms from the domain of video-games, cosmetics or hotel bookings. In this case, during the translation process, the system would probably first search for the term in the client's specific glossary, which would take precedence, and then in the general domain glossary. In this way, it would not be necessary to include domain-specific terms in all customer glossaries of a given domain, which would save time in glossary creation.

At Unbabel, interns were introduced to glossary creation, curation and translation as an integral part of their assigned tasks. All tasks related to the internship are explained in detail below.

2.3 Completed tasks and responsibilities

This section contains a chronological description of my assignments during the internship and the tasks I completed. There is also a brief description of some of the tools used to ensure the quality of the translations, a description of the structure of the glossary I worked on and the process of curating the glossary.

During the first part of the internship, following the suggestion of our supervisor at Unbabel, the four interns from the Quality team tested some of the company's tools and systems from the perspective of linguists and translators. This was done with the aim of providing feedback to the Product managers. The task was also to familiarise the interns with the company's translation workflow and help them carry out some of the steps in the workflow performed by humans, such

as post-editing of MT output, evaluation and annotation of translations, creation and translation of glossaries. This first part of the internship had a duration of 144 hours.

The second part of the internship started on 26 February 2018 and had a duration of 96 hours. It was adapted to the individual topic assigned to each intern and there were regular meetings with the supervisor and academic advisors. During this part, I worked individually on company glossaries and completed various tasks, which are described in more detail in section 2.3.5. To complete these tasks, I worked with different teams at Unbabel to gain more information about the functionality of the glossary system in the company's translation pipeline.

The supervisor provided the interns with the necessary hardware (personal computers to for use in the office), access to Unbabel's communication channels and to the translation software tools used, as well as additional documents and information about them. The interns were also supported by members of various technical and operational teams of the company.

We participated in an onboarding process that consisted of nine meetings with the leaders of each team in the company and a short presentation about the functions, tasks and objectives of each team; a short introduction of the interns through the company's communication channel and at the company's weekly general meeting.

After the onboarding, the interns performed the following tasks: testing some of the company's tools (Post-Editing Tool, Annotation Tool, Evaluation Tool); analysing the linguistic resources used at Unbabel (language guidelines, glossaries); discussing and preparing feedback on the tools and the resources; presenting this feedback at specially organised meetings; preparing five final reports. The final reports covered the following topics: Training Tasks Report - 3 pages, Language Guidelines Report - 6 pages, Glossary Report - 9 pages, Evaluation Tasks Report - 2 pages, Annotations Report - 3 pages. As these reports are only available in-house, please refer to section 2.3.4 for a summary of each report.

As far as my individual duties in curating the glossary are concerned, I was assigned the following tasks:

1. integrate the glossary into the company's new glossary template;
2. assign a part-of-speech tag to all source terms and revise them;
3. add the singular and plural forms for each term;
4. analyse the Portuguese and Bulgarian translations;
5. add gender-specific forms (feminine/masculine) where relevant;

6. check the spelling;

7. check the capitalisation.

Due to the size of the glossary (with 4224 source terms), the first three tasks were completed, as were the 6^{th} and 7^{th} tasks, but just for the source terms. The 4^{th} task was not carried out due to lack of time and because the analysis of the translations into Portuguese and Bulgarian would not yield sufficient results given the small number of target units in these languages. In the glossary, there were 582 target terms in European Portuguese and 0 target terms in Bulgarian. In comparison, there were 2481 target terms in Brazilian Portuguese.

2.3.1 Testing some of the tools of the company

We tested the three main tools related to quality assurance of translations produced in the company: the Post-Editing Tool, the Annotation Tool and the Evaluation Tool. These tools incorporate or are integrated with other tools, such as TM management systems, a spell-checking system and *Smartcheck*, which are also described below.

These tests are the result of performing real tasks in the production environment, which is the environment used by all professionals who work for the company.

2.3.1.1 Testing the Post-Editing Tool

We tested Unbabel's platform for **post-editing MT output (Editor's tool)**, which is used by thousands of bilinguals working remotely for the company in different language pairs. Each intern used her working languages. In all four cases, the source language was English and the target languages were Bulgarian (Nadezhda Metodieva), European Portuguese (Catarina Xavier) and Brazilian Portuguese (Tatiane Oliveira and Rhandra Lopes) depending on the mother tongue of each of us. After testing the platform and taking notes on possible problematic points regarding the functionality and use of the tool, the interns discussed them together and considered which points could be improved. A mind map was made on paper (following the method of Tony Buzan, 2006), and a presentation was prepared for the meeting with the members of the Product team. At this meeting we gave feedback on the features, design and difficulties we encountered while working on the platform. We made suggestions for improvements to the user experience and also from the perspective of professional translators, who have the challenges of the translation process in mind. This task was completed in two working days.

The Editor's Tool is used to improve the quality of MT output by manually correcting fluency, accuracy and style. These corrections are made by bilinguals, called *Editors*, who use a platform that displays the ST and the TT simultaneously and on which the latter can be edited. There are several other tools to help editors correct the text, such as glossaries and *Smartcheck*, described below.

First, we need to describe the roles of editors in the translation process. There are three types of **Editors**, according to Unbabel's official website (Unbabel, 2018e): Trainees, Editors and Senior Editors (Native Senior Editors), depending on their skills and also the experience they have in using the Editor's Tool. After registering on the website, the user has the status of Trainee. Only after completing ten training tasks and receiving a good mark in the evaluation process does the translator gain access to paid tasks and becomes an Editor. Editors are paid for post-editing MT outputs online. The number of tasks that an editor receives depends on several factors, including: volume of translations submitted by a client in a particular language pair; Quality and speed of delivery of previous tasks completed by the editor. Senior Editors are Editors who have received high ratings for the quality of previous tasks and have more experience working on the company's platform. They conduct Senior Reviews in the Editor's platform. Their work differs from regular post-editing of MT outputs in three ways: greater length of tasks, verification of the coherence of terminology used throughout a text and a direct dispatch of the task to the client. Senior Editors guarantee the quality of the final text as a whole and correct incoherent terminology, mistranslated or ungrammatical words or phrases. The native language of the Senior Editors is the language of the TT.

There are tools that help the editors and are part of the Editor's platform, such as: the *Smartcheck*, and the tools that integrate and search in the Glossaries and the Translation Memories (Unbabel, 2018b).

Smartcheck is a proprietary application developed by Unbabel that uses a set of rules to detect incorrect grammar, morphology, orthography and style in the TT. The aim of the application is not to automatically correct errors in the text, but to draw the attention of the Editors to problematic parts so that they can correct them. These parts of the text appear underlined with a red line. The system indicates the number of problems that need to be corrected and ensures that the Editor does not submit a task without solving them first. One way of correcting an issue is to replace the problematic unit with another one that is recognised by the system. Another way is to accept the underlined unit as correct and add it to a list of units approved by the users. At the time the internship took place, *Smartcheck* did not work properly in language pairs with a low volume of translations submitted by clients, such as

English – Bulgarian, so observations in these language pairs were sparse. The same applies to the glossaries and translation memories.

In the Editor's tool, **glossary** words are highlighted both in the ST and in the TT (see Figure 2.2[4]). Sometimes it is also possible for the Editors to edit them (e.g. inflect them in gender and number, if necessary). However, according to the instructions provided in Unbabel official website (Unbabel, 2018d) it is not desirable to replace glossary terms. This should only be done in special cases, for example if the target term is inappropriate in a particular context.

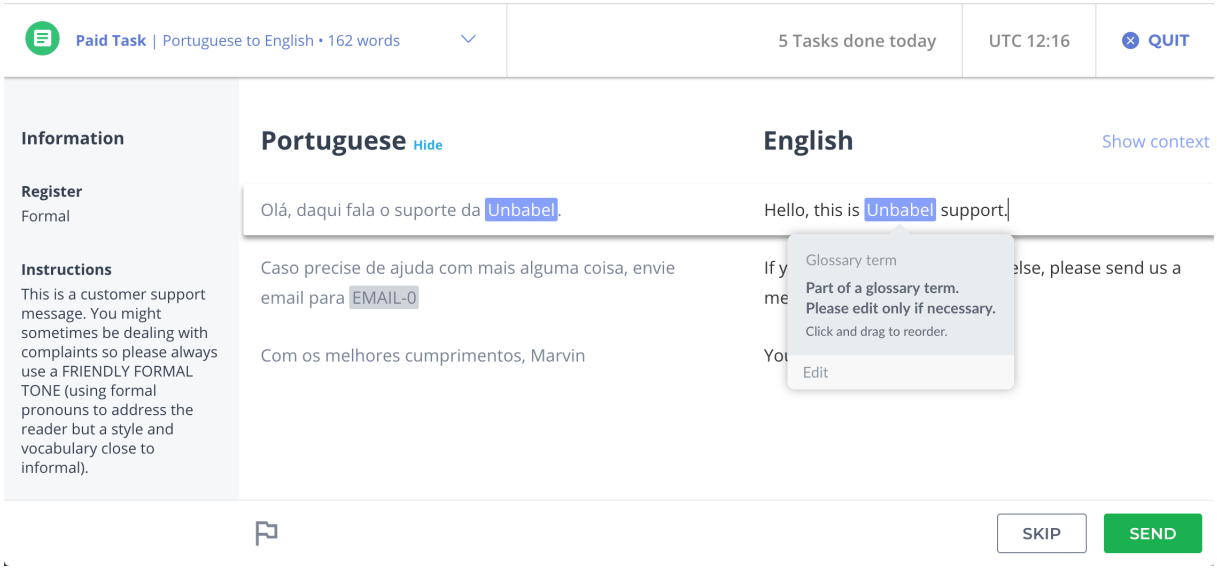

Figure 2.2: A screenshot of the Post-Editing Tool with highlighted glossary words.

Translation Memories store segments of text that have already been approved by human translators. These are used as suggestions for new translations when a segment of the ST matches a stored segment exactly. TMs help to ensure consistency between translations of different texts submitted by a particular client and to deliver the new texts faster. Segments marked as "Translation memory" should only be edited when absolutely necessary, as modifying them affects the stored segments and makes it more difficult for the system to deliver the best match to the stored texts. However, if an Editor discovers an error in a TM segment, it is highly recommended that he corrects it.

It is important to note the difference between training tasks and paid tasks of post-editing MT outputs. Although the interface of the Post-Editing tool is identical for both types of tasks and both must be completed within a certain time frame, training tasks are created in advance by Unbabel by selecting source and target texts and deliberately modifying a well-translated

[4]Reproduced from the information available at the Unbabel website: https://help.unbabel.com/hc/en-us/articles/360003366273-What-are-Glossaries-

TT and inserting some incorrect units that need correction. Paid tasks, on the other hand, are source texts submitted by a client, machine-translated at Unbabel and sent to a post-editor. This indicates that training tasks are limited in number and variety, and in general they need post-editing. In contrast, paid tasks correspond to an unlimited number of source texts where, depending on many factors such as the performance of the MT system and the presence of segments previously saved as TMs, corrections may not be necessary.

The interns tested the Post-Editing tool as Trainees. The final reports therefore made some suggestions for improvements to the user-experience on the platform, as well as suggestions related to the training tasks.

2.3.1.2 Testing the Annotation Tool

In order to test the Annotation Tool and become familiar with this type of task, translation annotation tasks were completed in the **Annotator's Tool** (one batch of 30 texts per intern, which took four working days to complete). These tasks were completed after careful reading of the Annotation Guidelines[5] and the Language Guidelines[6] for each intern's working languages (English, Portuguese and, in my case, also Bulgarian). Once this task was completed, another presentation was prepared for the Product team, which included feedback on the annotation platform and comments on the Language Guidelines.

The Annotation tool is a tool developed at Unbabel and used by a team of professional linguists (called **Annotators**) who work remotely for the company, annotating errors in translated texts in many languages. According to Comparin and Mendes (2017b), given the variable quality of MT output, error annotation is suitable "to evaluate the quality of the results produced by a MT system, but also to outline strategies to improve them and reduce the number of errors in the output produced". At Unbabel, quality audits are conducted regularly "to be able to objectively compare our [Unbabel's] performance with third parties and open source translation" (Unbabel, 2017a).

The company has based its Annotation tool and annotation guidelines on the Multidimensional Quality Metrics (MQM) framework (Comparin and Mendes, 2017a), which is used in the industry to assess the quality of translated texts and to identify specific issues in those texts (Lommel et al., 2014). This framework does not promote individual metrics, but is intended

[5]Internal document, not publicly available.
[6]Documents available only to registered users of the official Unbabel website.

to be used for all translations so that "quality can be defined by how well a text meets its communicative purpose" (DFKI, 2014).

The MQM measures the quality of a translated text in different aspects, for example, accuracy, fluency, style, etc. There is a detailed decision tree used in the company, as well as some clarifications in the Annotation Guidelines which help annotators make decisions about problems in the texts. There are also three levels of severity of the issues: minor, major and critical, which are weighted differently and are fundamental to the attribution of a Quality Translation 21 (QT21)[7] score to each annotated text (DFKI, 2014).

The QT21 is a numerical score assigned to the annotated translations at Unbabel. It is expressed using a special formula, which takes into account the number of translation problems in a TT and their severity, as well as the total number of words in the annotated text. The initial score is 0 and the higher the score, the lower the quality of the translated text.

The types of errors used in the Annotation tool of Unbabel and the penalty system used to assign a numerical value to each error are clearly described in Comparin and Mendes (2017b) along with the main categories of errors: "accuracy, fluency, style, terminology, wrong language variety, named entities, and formatting and encoding". The analysis of these errors, as mentioned by the authors (*ibid.*), can not only make visible which types of errors occur in machine-translated texts and in human post-edited texts, but also show the significant quantitative reduction of errors after human post-editing of machine translation.

For some of the language pairs there are several annotators. Since the QT21 score is assigned to all annotated translations, it is necessary to ensure that there is agreement between the annotators in the annotation of errors. This agreement is mainly achieved through the Annotation Guidelines, but good practice might be for annotators to discuss the types of errors and establish some common understanding of the examples in the guidelines and some specific ways of working for their language pairs.

The texts I was given during the internship to annotate in the language pair English – Bulgarian were very short and some of them consisted of a single word in the ST, which made the task very challenging due to the lack of context.

In order to have texts annotated, the Quality team assigns an annotation batch to an Annotator. After accessing the Annotation tool platform, the Annotator can see a page with all the

[7]QT21 is a machine translation project, funded by the European Union's program "Horizon 2020" which aims to "substantially improve statistical and machine-learning based translation models for challenging languages and resource scenarios", as well as to improve the evaluation and the analysis of quality barriers, etc. (QT21 Consortium, 2018). Parts of the MQM framework were developed within the QT21 project.

annotation batches, assigned to him/her and their status. Once an annotation batch is opened, the first of the annotation tasks included in it is displayed automatically on the screen. The interface for each task includes a ST and a TT; the number of tasks in the batch; specific instructions from the client (if any); information about the register to be used (formal or informal); highlighted words corresponding to glossary terms in the ST and TT (if any); and an annotation control panel. The TT can be one of these two types: a MT output; or a MT output that has been post-edited by one or two Editors and, in some cases, also by a Senior Editor.

The annotation process is performed in the following way: the Annotator identifies an error in the TT and selects the incorrect word or combination of words. After that he has to select the type of error and its severity. The error types are divided into three main categories, such as fluency, accuracy and style, which include numerous subcategories. For each subcategory that can be selected, there is a definition in the Annotation guidelines [8] provided to each Annotator. Once the type of error is selected, the severity level must be determined. The three severity levels are minor, major and critical, depending on how a particular error affects the readability or the accuracy of the TT.

Annotated errors appear highlighted in the text. Once all of them are annotated and the task is ready to be submitted, the Annotator has to assign a "fluency level" to the TT, which can vary between 1 and 5 stars. The higher the number, the more "natural" the translation is in the TL, according to the Annotation guidelines.

The functionality of this tool varies greatly from the Evaluation tool, described in the next section.

2.3.1.3 Testing the Evaluation Tool

After the Annotation task at the internship, an assignment to **evaluate editors** was completed on a special platform used by Unbabel's Evaluators. The materials provided to the interns for this task were the Evaluation Guidelines (common to all language pairs)[9] and access to the Evaluation tool. The interns evaluated various translations in their working languages and prepared comments and suggestions for improvement of the platform and the evaluations which were presented to the Product Team. The task was completed in two working days.

Evaluators check the quality of the post-editing tasks submitted in the post-editing platform. According to Unbabel's official website (Unbabel, 2018a), the evaluations are "unbiased, objective

[8] Internal document, not available publicly
[9] Internal document, not available publicly.

quality checks" that are not optional. They are also unpredictable and randomly selected from the tasks completed by a particular post-editor. The role of the Evaluator is to decide whether a translation task complies with the Unbabel style guide (i.e. the Language Guidelines for each language available on the official website) and whether the translation accurately reflects the meaning of the ST. As a rule, the Evaluators are linguists or professional translators who evaluate target texts (TTs) in their mother tongue.

Since Unbabel's post-editors of MT output are usually bilinguals with no translation experience, the aim of the evaluation is to select those among the Trainees who meet the requirements for working for the company in the language pair that is being tested. Some of these requirements are: being a native speaker of the TL; having the ability to translate the meaning of the ST accurately and precisely; having knowledge of the grammar and orthographic rules of the TL; being able to complete the tasks in the given time frame. Evaluations are also used to ensure that the expected quality of post-editing is achieved throughout the paid post-editing tasks that the Editors submit.

Therefore, there are two types of evaluations that can be made in the Evaluation tool, namely evaluations of tasks post-edited by Trainees or by Editors. These tasks are randomly selected by the system, sent to the Evaluation pipeline and assigned to an Evaluator working in that language pair. When an evaluation task is completed, it is sent back to the evaluated post-editor and the records for each evaluated task are also available to the Evaluator. Neither the Evaluator nor the Editor has access to each other's identity. This is a measure to prevent biassed and/or fraudulent evaluations.

Trainees are evaluated upon completion of 10 training tasks and Editors are evaluated on a regular basis. Evaluators must assign a score between one and five stars to each evaluation, with the five stars score being the highest possible score and corresponding to a good quality translation. Once five translation tasks have been evaluated, the average score of those is calculated.

Trainees can change their status to Editors and take on paid translation tasks only if their average score is above a certain threshold. As mentioned above, the quality of translations provided by the Editors is also checked through the Evaluation process. This means that a Trainee who, for example, achieves a good average score and is promoted to Editor, but later delivers translations that are below the threshold, is demoted to Trainee.

The Evaluation Guidelines define the rules for the evaluation process and provide recommendations for the attribution of the score, according to the severity and type of the errors.

For example, minor orthographic errors do not have much impact on the score, but if there is a major error that changes the meaning of the text, the score cannot be higher than one star, which would undoubtedly affect the average score of the post-editor. The score is also used to determine the hourly pay rate of an Editor. The higher the score and the less the time spent on post-editing – the higher the hourly pay rate.

The Evaluation process is carried out as follows. On the Evaluation platform, the Evaluator can see both the ST and the post-edited MT output, together with the highlighted glossary words. There is also the option to see either the final version of the post-edited MT output, or the first version of the MT output provided to the post-editor and the corrections made by him/her. This option allows the Evaluator to detect whether or not a particular post-editor has made corrections to the MT output and what kind of corrections have been made. Sometimes, the quality of the MT output is acceptable and the post-editors can submit it without corrections. However, often the quality needs improvement, but the post-editors submit it without corrections. One of the main purposes of the evaluation is to find out which types of corrections are a priority for a particular post-editor and which are not done, either because they were not identified, because they were considered irrelevant or unnecessary, or because there was not enough time for post-editing.

In the Evaluation platform there is also the score assessment panel, which consists of a comment area and a score area. Evaluators must write a comment regarding the overall quality of the translation and on the units that were translated in an incorrect or imprecise way and make suggestions for their correction. Therefore, the aim of the evaluation comment is to justify the score assigned and to explain the errors in detail, as in some cases post-editors cannot identify them on their own. The errors identified and comments are, of course, in line with the Evaluation Guidelines and the Language Guidelines of each target language, mentioned earlier.

The comments on the Evaluation platform are written exclusively in English. This allows the relevant teams at Unbabel to understand and analyse both the performance of a particular Editor and the performance of the Evaluator.

At the beginning of the internship, the Evaluators had to provide a comment on each evaluation, which takes some time if there are many errors. Given the fixed number of training tasks (see p. 19) and the fact that some post-editors submit the MT outputs without editing them, the interns did not consider this practice of leaving a detailed comment very efficient for two reasons. The first is that the comments offered easy "solutions" to some difficulties regarding the correct translation of certain elements of the ST, which could be used fraudulently to achieve a

good average score and become an Editor. The second reason is that there is no point in wasting the Evaluator's time and the company's resources evaluating a TT that is identical to the MT output. Therefore, one of our suggestions was to consider the possibility of not writing a comment for certain evaluations. This suggestion has been implemented in an updated version of the Evaluation tool, where Trainees' evaluations do not include a comment section, while Editors' evaluations do.

All tasks described above, i.e. testing the Post-Editing tool, the Annotation tool and the Evaluation tool, were dependent on the linguistic resources available in the company. These are described below in the context of a task dedicated to their analysis and improvement.

2.3.2 Analysis of linguistic resources

Some of the linguistic resources that are created and used in the company are the Language Guidelines, which are specific to each target language, and the Glossaries, described in section 2.2.2. These are included here because they are related to the completion of the tasks mentioned.

2.3.2.1 Language Guidelines

As part of the preparation for fulfilling the tasks described in section 2.3.1, the interns had to analyse the Language Guidelines (Unbabel, 2017b) for their mother tongues and suggest improvements. The Language Guidelines are the official style guides for each target language used in the company. The guidelines consist of language-specific grammar and punctuation rules, as well as some suggestions for the post-editing of MT and the use of formal and informal registers. They follow a common structure for all languages in order to standardise as much as possible the type of information provided and to make it easier for post-editors to find information on a particular topic.

The Language Guidelines for all languages could be accessed freely on the dedicated section of the Unbabel's website (Unbabel, 2017b).

In my case, I considered the Language Guidelines for Bulgarian. I checked the grammatical and punctuation rules described in the text and provided comments during the meeting with the Product team.

These guidelines serve as the basis for testing and using the tools described above and are fundamental to the agreement between all parties involved such as post-editors, evaluators, annotators, the company and its clients.

While the Language Guidelines are an important aspect of post-editing, annotation and evaluation at Unbabel, they do not fall within the main scope of this report. Therefore, I will not elaborate on the topic in this document, as it goes beyond its focus.

2.3.2.2 Glossaries

As for the glossaries, the interns were asked to analyse the multilingual glossaries, focusing on the selection and translation of terms. As before, each intern had to compare the English version of the glossary with the translation of the terms into their mother tongue or working languages. The default source language of all glossaries is English. This means that the selection of terms to be included in the glossary is done in English and then translated into other languages selected by the client.

In preparation for this task, a meeting was organised with the person responsible for glossary creation and implementation at Unbabel, where she provided important information about the glossary system and its use. Due to the sensitivity of the system and the lack of a dedicated platform for testing, the interns were not given access to the glossary platform. Instead, they were provided with various XLSX files containing the information available on the glossary platform.

In addition to assessing the quality of the glossaries used at Unbabel for specific clients, the aim of this task was to combine the information encoded in different glossaries for the same client into a single glossary in order to eliminate unnecessary units. The task consisted also in understanding how the glossary system works and identifying the problematic points.

Due to the volume of terms provided, which included about 2,000 source terms, the task could not be fully completed, but the results were sufficient to make important observations about glossary creation and translation and to describe them in detail at the meeting with the supervisor and the Product team. The result was also compared with a new version created by the team responsible for the glossaries using the same method and the differences were analysed by the interns.

The glossary provided did not belong to the same client as the one analysed in Chapter 5 of this report, although the business field of these clients is the same, namely video games. Furthermore, it is important to mention that this assignment aimed to give all interns a general overview of the topic. The detailed analysis of a similar glossary was carried out in the second part of the internship, which was dedicated to the individual interests of each intern (see section 2.3.5).

2.3.3 Preparation of feedback and presentations

The company, more specifically the Product team, was interested in the feedback that the interns could provide on the tools, resources and workflows tested. The team felt it was important to understand what points could be improved from the translators' perspective and also to improve the user experience on the platforms.

Thus, as the interns had the same workspace, it was easy to discuss the tools and share observations and comments during the testing.

The method used to prepare for the meetings with the Product team was the Mind Map method (Buzan & Buzan, 2006), which combines brainstorming with schematic representation of ideas and was very useful for the efficient discussion of ideas between the interns and with the Product team.

Also, regular meetings were scheduled with the supervisor and the academic advisors where the interns received guidance and resources regarding their internship reports and their preparation.

2.3.4 Elaboration of Final reports

This subsection presents some brief conclusions from the five final reports prepared by the interns in relation to the tools tested and the resources analysed. The task consisted in the elaboration of short texts on the tasks mentioned earlier in subsections 2.3.1 and 2.3.2. The preparation of the texts was divided among the interns as follows: each one of them was responsible for writing the main part of one report and the final revision was done by all of them, so that the suggestions for editing were shared.

In the **Post-editing Tool** (Editor's tool) some problems were found, such as: non-functioning glossaries in some language pairs; training tasks use a small number of texts, resulting in repetitions of the same texts; unclear user interface regarding the time frame for post-editing a text; confusing client instructions and lack of context of the specific segment being edited. In addition, due to the anonymisation of client names, the gender marker in the English ST was lost, leading editors to assign a masculine gender in all texts translated into languages with grammatical gender. Specific improvements to the user interface were suggested, as well as an alternative anonymisation method, a more diverse selection of training tasks and more informative contextual elements.

Regarding the **Annotation Tool**, some of the issues identified were related to: the Annotation Guidelines provided for training Annotators; the content to be translated (e.g. content that needed to be localised by a professional translator was submitted for MT and post-editing without sufficient context); the lack of contextual information about the client; and to some technical issues related to the glossaries or their reliability.

The suggestions, to name a few, were: improvements in the Annotation Guidelines and elaboration of different guidelines for different language families; improvements in the tool interface (visibility of the text and the severity of errors, indicated by appropriate colours); and the inclusion of a "Go back" button and the possibility to put tasks on pause and work on another one in the meantime.

The problms affecting the **Evaluation Tool** were similar to the ones in the other tools tested (e.g. lack of context, register and glossary words not highlighted). In addition, some comments were made about the accuracy of the evaluation score. The suggestions included corrections in the Evaluation Guidelines, a more precise system for error evaluation (similar to that in the Annotation Tool) and the possibility of not writing comments if the task has not been post-edited at all.

With regard to the **Language Guidelines**, detailed suggestions for improvement were made on the problems identified, relating to its structure, grammar, style and orthography, its language variety and the examples included in the text. A suggestion was also made for easier access to the Language guidelines.

The report dealing with the issues in the **Glossaries** mentions some general problems, such as lack of instructions (e.g. there were no guidelines for creating or translating glossaries), lack of descriptions of the terms, missing target terms, inconsistencies regarding the choice of plural or singular form, as well as some specific problems of the glossary analysed, along with suggestions to improve some incoherent, incorrect or inappropriate translations and the exclusion of some source terms.

The interns were able to observe how, over time, their suggestions at the meetings and in the final reports became new features in the tested tools, thanks to the different teams in the company. For example, a meeting with the Product team that took place on October 3rd, 2017, to discuss the interns' suggestions on the Editor's Tool resulted in a new look and new functionalities in the tool, for the interns to test on November 5, 2017. Some of the interns' suggestions on the Language Guidelines were also included in the new versions of these documents, which are available online. Glossary Creation Guidelines were elaborated quickly afterwards. The

improvements in the Annotation Tool, some of which were suggested by the interns (e.g. the possibility to pause a task and take another) and the improvements in the Evaluation tool (e.g. the possibility not to write a comment on every evaluated task) were incorporated into the new versions of the tool launched in the first semester of 2018.

2.3.5 Glossary Curation

The glossary I worked on was created for a client in the online video game industry. The client's name and other personal identifiable information are not mentioned in this work, in accordance with the legal restrictions related to the General Data Protection Regulation (European Parliament & Council of the European Union, 2016), applicable since May 24, 2018 in the European Union (see section 4.3).

According to the information provided by the Operations team, the intended use of the glossary was MT and post-editing the communication between Customer Support Service (CSS) agents and users of the video games.

Three versions of the same glossary were provided (referred to here as **G0**, **G1** and **G2**). All contained source terms in English and were in the offline XLSX file format extracted from the online glossary platform.

The work on G0 and G1 was carried out during the period of the internship. G2 was considered in the elaboration of this internship report. For this reason, its analysis is presented in Chapter 5.

The G0 glossary referred to here was actually a set of glossaries representing each of the client's brands (a single video game), and the associated task was to integrate all glossaries into one.

The client provided the following information – an in-house glossary, used for the localisation of the games (see section 4.4), short instructions, an official website and a customer support page translated into the languages of the localised versions of the video games.

My assignment for this part of the internship included the following tasks (as already mentioned in section 2.3) which correspond to the versions of the glossary, as shown below:

G0:

- integrate the glossary into the company's new glossary template;

G1:

- assign a part-of-speech tag to all source terms and revise them;
- add the singular and plural forms for each source term;
- check the spelling of the source terms;
- check the capitalisation of the source terms.

G2:

- analyse the Portuguese and Bulgarian translations;
- add gender-specific forms (feminine/masculine) where relevant;

G1 had very few target terms in my working languages (Bulgarian and European Portuguese). Therefore, the analysis was focused on the source terms - selection, description, etc. In relation to G2, there I focused on the translations into European Portuguese and Bulgarian. The target terms in Brazilian Portuguese are included for comparison.

G1 had a total of 4224 source terms and was created from a **glossary provided by the client** (see section 4.4) in combination with **terms selected by Unbabel** and words that were **automatically extracted** from text samples provided by the client (emails and other customer support communications).

It is important to note that G0 and G1 were working versions of the glossary. In the final version (G2), the team responsible for creating and managing the glossary at Unbabel, excluded about 19 % (802 entries) of the terms included in the G1.

2.3.5.1 Integration into the new glossary template

The initial version of the XLSX file (G0) of the glossary contained several sheets (more than 30), each containing the units corresponding to the different brands (i.e. the different video games) of the client. There was also a sheet with the new template for glossaries at Unbabel, another with instructions on how to complete the task, and an "All" sheet, that listed all the client's brands terms, merged together.

I completed the integration of the glossary into the new glossary template. However, due to the dynamic creation, translation and curation of glossaries in the company, this first working

version was not used for the next tasks. Then, another, a slightly more complete version was provided to me (referred to here as G1).

Nonetheless, in order to understand the changes in the glossary system, I think it is important to explain the original structure of the glossary (G0) as shown in Table 2.1, which was later replaced by the current structure of the glossaries at Unbabel.

Table 2.1: Initial structure of the glossary, extracted directly from the platform.

id	[SOURCE UNIT]	description	frequency	[TARGET UNIT]

In the G0 version of the glossary mentioned above, the column *id* was always empty; the *[SOURCE UNIT]* column listed the units in English, one per line; the *description* column contained brief definitions of the units; the *frequency* column was designed to encode the frequency of occurrence of the units in the client's documentation. It was filled in with a value of *1* for all units, so it is irrelevant for this analysis. The last field was the *[TARGET UNIT]*, which contained separate columns for each target language. However, most of them were incomplete and contained only a very small number of translated terms.

The new glossary template, shown in Table 2.2 and Table 2.3, includes columns for additional information. These have been included to facilitate the work of the glossary translators and to make the translation process faster and more accurate by providing additional information on the units.

Table 2.2: Information regarding source units in the updated glossary template

[SOURCE UNIT]	Origin	Part of Speech	Translate?	Description	Frequency

The column *Origin* contained the origin of the source unit, viz: curated materials provided by the client, such as glossaries (marked as *Client*); Unbabel research in the client's official website or user forums (marked as *Unbabel*); units that had already been included in the glossary platform (marked as *Platform*); units extracted from incoming or outgoing customer support emails and other communication of the client (marked as *Tickets*).

G1 had a total of 4224 source units, 770 of which (18.23 %) came from the in-house glossary of the client, 2009 (47.56 %) were taken from the glossary already created on the Unbabel glossary platform and the remaining 1447 (34.26 %) were automatically extracted from tickets, as shown in Figure 2.3.

The information provided in this column is very important because the origin of the unit can help the creator of the glossary assign a reliability value to the candidate units during the

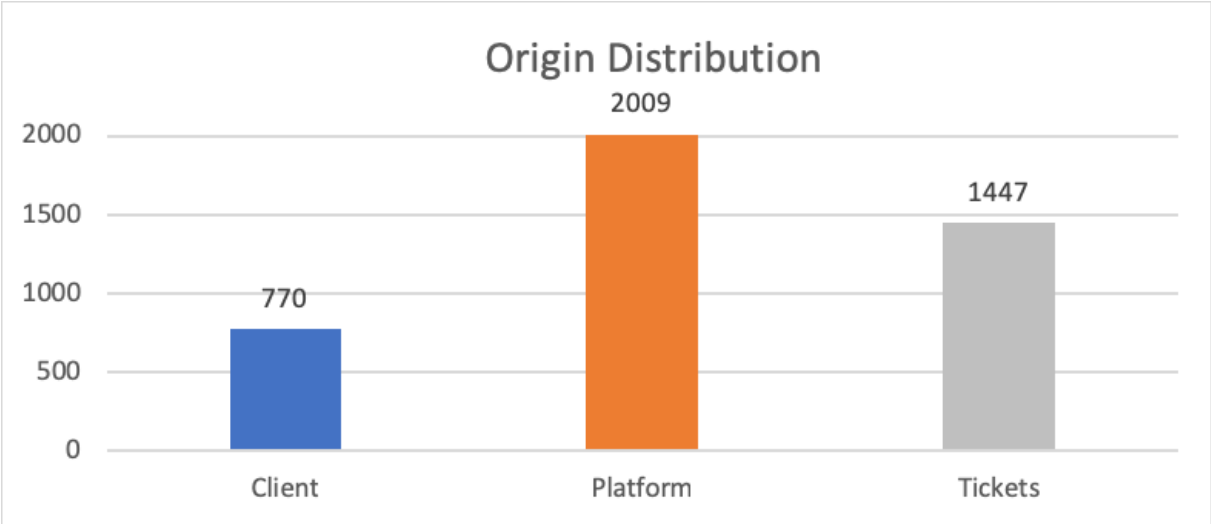

Figure 2.3: Distribution of source units, according to their origin

selection and curation of the units.

On the one hand, units provided by the client are likely to have a greater chance of being selected for Unbabel's glossary, as they probably have already been revised by terminologists or in-house translators. In addition, the units that come from Unbabel research are usually selected to meet the specific requirements of the system processing the glossaries and the requirements of the post-editing of the MT output. These include, for example, the units that should be kept in their original form (e.g. names of products, trademarks and others). This column also serves as a reference for the translators when searching for additional context.

On the other hand, the units that come from customer tickets (e.g. emails and messages) are usually automatically extracted from samples of the customer's communication channels. It is important to note that many of the entries extracted in this way had various different problems and did not comply with the glossary guidelines presented in section 2.2.2.2, and therefore had to be excluded. For example, they contained parts of sentences that had been inadvertently selected because of their frequent occurrence in the sample texts. These include ungrammatical constructions (e.g. *amount money*), parts of the next sentence and structures which, taken out of the overall context of occurrence, may have different syntactic relations between elements.

Cases like these highlight the limitations of the automatic extraction system, and lead us to consider the extracted terms as less reliable. In fact, only a very small number of them were included in the revised glossary (G2): only 160 entries out of the 1447 originally listed in G1.

Nevertheless, even if the units were taken from glossaries provided by the client, they had to be carefully revised. In some cases they contain obsolete words and expressions, that are no

longer used by the client, or they are missing new words that came into use after the company's glossary was created and were not added afterwards. For this reason, the automatic extraction method is used. In summary, Unbabel is also not guided exclusively by the client's resources.

The inclusion of the *Part of Speech* column was motivated by the need to disambiguate homographic units in the glossary. For example, to distinguish verb forms from noun forms in the source language and thus avoid incorrect or incoherent translations.

The column *Translate?* has been included to inform glossary translators whether a particular unit needs to be translated or not. For example, some clients require that the names of brands or certain products be kept in their original form. The possible values for this column were restricted to *YES* or *NO*. In G1, however, it was empty.

In the column *Definition* many definitions were composed by a short explanation, showing only the category of the unit in the game: e.g. *Decoration*; *Helper*; *Episode*, etc. Considering the large number of units, many of these definitions were repeated several times and did not make the meaning or characteristics of the unit described any clearer.

In G1, in the *Frequency* column, 77 % of the values were filled in. Those whose origin was *Platform* or *Client* always had the value *1*, while those originating from *Tickets* had different values. However, the highest frequency values were observed for common words such as *access*, which should not be included in the glossary.

Table 2.3 shows a model repeated for each of the target languages. The information in the new columns is provided by the translators for each language. These were designed to be used by the team responsible for implementing the glossaries and by the system applying the glossary to the target texts to facilitate post-editing.

Table 2.3: Information regarding target units in the updated glossary template

[TARGET UNIT]	Enforce casing?	Translation Reference	Invariable?

The columns for the target units shown in Table 2.3, which were multiplied 26 times, one for each target language, as previously referred, were incomplete or empty. This is understandable because at this point the new structure of the glossary was being defined in order to be sent to the translators of each target language afterwards.

The column *Enforce casing?* was designed to indicate whether or not the capitalisation of the glossary entry should be maintained. For example, if it is set to *YES*, the system locks the capitalisation of the target unit and it is applied directly to the text for post-editing. If it is set to *NO*, the capitalisation is unlocked and the unit appears in the TT with the same capitalisation

as in the ST. In some cases this function is very useful. For example, to reflect the original capitalisation of brands, names and so on. In this way, the specific units of the glossary already appear capitalised in the TT and the editor does not have to edit them.

The column *Translation reference* is used in the revision of the glossary by the team responsible for the quality of the translations and the glossaries. In this column, the translator can justify his or her choice of translation for a particular source term.

The column *Invariable?* is used to "lock" the unit so that editors cannot change it. This is used in case the target term would not appear in any other form in the TTs. In other words: if the target unit may have different inflected forms depending on the context, this option should be set to *NO* so that the editors can adapt it to the linguistic context in which it occurs. This option is also useful for product or brand names, which should not vary in form.

2.3.5.2 Part-of-speech tagging

As mentioned earlier, one of the tasks assigned to me consisted in manually tagging all source terms with their PoS tags. This task was very time-consuming due to the size of the glossary I was working on.

I was instructed to fill in the blank spaces with PoS tags indicating the word class of the unit. The tag set was inspired by the one found in the Universal dependencies project (Universal Dependencies & Nivre, 2017) and included 17 categories, as shown in Table 2.4.

Table 2.4: PoS tag set used at Unbabel for tagging glossary units

Word class	Tag	Word class	Tag
adjective	ADJ	particle	PART
adposition	ADP	pronoun	PRON
adverb	ADV	proper noun	PROPN
auxiliary	AUX	punctuation	PUNCT
coordinating conjunction	CCONJ	subordinating conjunction	SCONJ
determiner	DET	symbol	SYM
interjection	INTJ	verb	VERB
noun	NOUN	other	X
numeral	NUM		

As predicted by terminology theory, the majority of the terms in the glossary consisted of noun-based structures.[10] And this was indeed confirmed after completing the task.

[10]As mentioned in Cabré and Sager, 1999, p.112: "Nouns represent two-thirds of all terms. One of the peculiarities of specialized terminology is the tendency to prefer noun-based structures, as opposed to verb-based or

The results of the manual tagging can be observed in Figure 2.4.

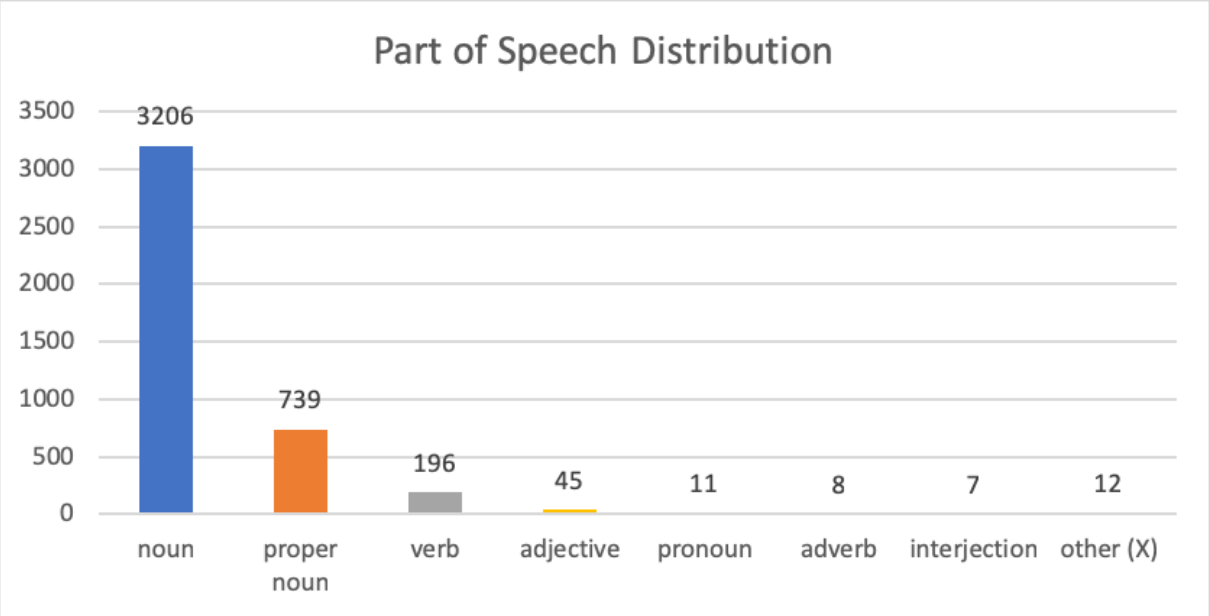

Figure 2.4: Part of Speech distribution in G1

As can be seen in Figure 2.4, of the mentioned tag set of 17 categories, only 8 were represented in the glossary: *noun, proper noun, verb, adjective, adverb, pronoun, interjection, other*. The most numerous category was *noun* – 3206 units (representing 75.90 % of the total units), followed by *proper noun* – 739 units (17.50 %), *verb* – 197 units (4.64 %) and *adjective* – 45 units (1.07 %). The other categories such as *adverb*, *pronoun* and *interjection* had less than 15 units each (0.62 % in total). The last category, marked with *X* (*other*), represents units that could not be included in any of the other 16 categories, such as URL addresses and abbreviations.

Units that had the origin *Client* included terms from the categories *noun* (729 units), *proper noun* (38 units) and *verb* (3 units). The presence of the categories *adverb*, *pronoun*, *interjection*, and *other* in the glossary is thus mainly due to automatic term extraction or manual term selection, as the Figures 2.5 and 2.6 show.

The glossary we are working with is intended for human and machine use. So, considering that the terms will be applied before the MT and will have an impact on the MT and the post-editing, the presence of function words like pronouns in the glossary is surprising. Another problematic point is the inclusion of words that have a wide variety of inflected forms in some languages. In a glossary intended exclusively for human use, the presence of verbs is accepted, but in our case verbs could only be included in exceptional cases.

adjective-based structures, to designate the concepts in a subject field. The cells in the conceptual structures of a domain take the linguistic form of nouns."

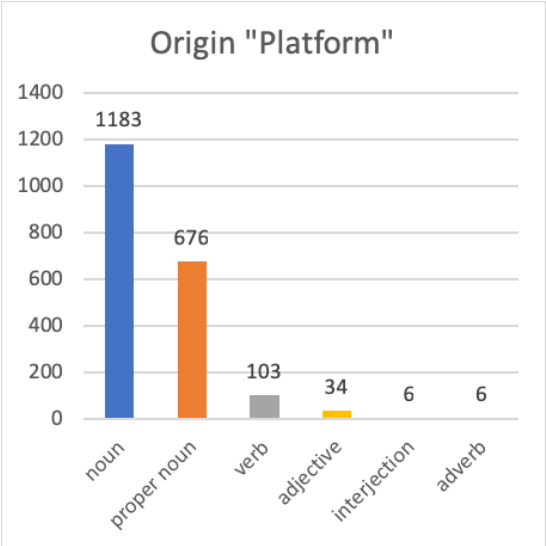

Figure 2.5: Distribution of PoS tags for terms with origin *Platform*

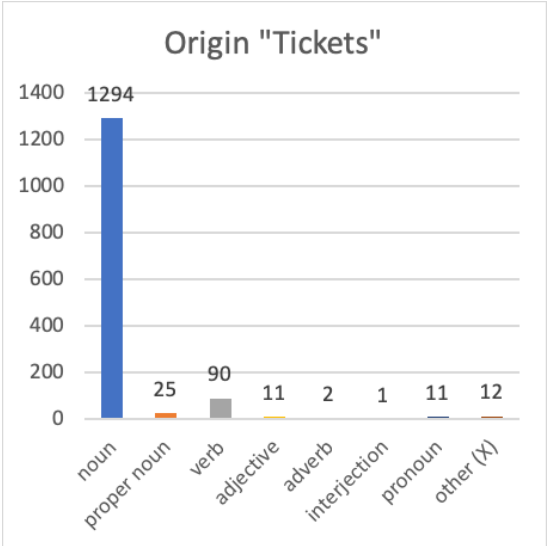

Figure 2.6: Distribution of PoS tags for terms with origin *Tickets*

2.3.5.3 Adding inflected forms in number

One of the tasks consisted in adding plural or singular forms to some of the terms because, as already mentioned, the glossary system cannot distinguish between the different inflected forms of the words. Therefore, if we consider a term that appears only in the singular form in the glossary, but appears in the plural in a ST, the system cannot recognise it. In this case, the glossary becomes ineffective.

For this purpose, I created two additional columns in the source terms section called *singular* and *plural* to encode the plural and/or singular forms of the entries: 735 plural and 146 singular forms of nouns were added. In order to make this information usable for automatic systems, the final goal of this task was to automatically split the entries at a later stage using the information manually encoded as described.

In carrying out this task, I was also able to find out how many terms had already been split, i.e. for which the glossary contained both the singular and the plural form. There were 142 such terms.

In some cases, it was not necessary to include the plural form of the units in the glossary because they are never used in the plural. For example, not all names of episodes, characters or objects in the video games exist in the plural form; however, there is no way to find out which ones unless one plays the video games.

2.3.5.4 Filling in the column *Translate*

This task was not originally assigned to me, but I decided to undertake it with the intention of finding different patterns in the glossary units and to facilitate the tasks of term selection and translation. As mentioned in the previous section, the cells in the column *Translate* could only be filled in with *YES* or *NO*.

Since the domain was video games and some of them did have an official translation in some of the target languages, several challenges arose in terms of translation/non-translation of in-game elements, such as: names of buttons, characters, episodes; information that appears in the game; instructions on how to solve problems, etc.

For example, let us imagine that a Portuguese user is playing the English version of a game because there is no localised version in Portuguese. This player then encounters a problem and tries to communicate with a CSS agent in his or her native language. However, if this person translates the names of buttons, characters and other in-game elements into Portuguese, it may be more difficult for him/her to be understood by the agent trying to help.

It is therefore necessary to summarise once again the steps involved in translating incoming and outgoing CSS communication at Unbabel. In general, a message written by a user in one of the SL available at Unbabel is received for translation; it is translated into English, the working language of the CSS agents; then it is read by the CSS agent, who responds in English and sends the text back to Unbabel to be translated into the required TL and then sent back to the user.

So if a user (a player of the game) uses a glossary word, but translates it, the glossary processing system will not recognise that particular glossary unit and the post-editors will not know that it is a glossary word. Furthermore, if the post-editors decide to translate it, there might be problems translating elements that require creativity.

The reverse is also possible: consider the case where an agent uses a version of the in-game elements translated at Unbabel and not an official localised version, in the absence of one. Therefore, the new names of the elements will lead to strangeness and probably unsuccessful communication, as the user would not be able to decipher the message of the CSS agent.

Still, this cannot always be predicted by the glossary creators, so additional information should be requested from the client about these special cases. Or instead, Unbabel can access the context of the glossary words and make a decision about whether or not to translate certain glossary words by building and analysing a corpus of previous translations for the same client.

However, based on my experience with the glossary and at stage in the development of

the glossary system considered in this work, I believe it is important to keep the names of episodes and some other in-game elements in their original form to make it easier for customers to communicate with the CSS agents. In many cases, for example, players need assistance with a certain situation in the game and the CSS agents provide instructions on how to solve the problem. Therefore, these instructions must be well understood and followed.

Furthermore, during the execution of the task, I found that in some cases the option of translating or not a particular term in full was not suitable. For example, when part of the term had to be translated and the other part had to be left in the original, e.g. in *smartphone /*BRAND*/*. For the first word there is a corresponding equivalent in each target language and the second must be kept in the original due to copyright or other legal restrictions. For this reason, a new value has been introduced: *MISC* for *miscellaneous* and 114 terms were marked with it.

In addition, a separate column called *Transliterate* was created to define which of the terms need to be transliterated or transcribed phonetically if the target language uses a different writing system of the one used in the SL.

My first intention was to use it to define which names of companies or products can also be transliterated. For example, languages that use the Cyrillic script sometimes use the phonetically transcribed version of the company name *Google*, which in Bulgarian is *Гугъл* and in Russian is *Гугл*. However the practice in the company is to leave all of them in the original script which is also done with other brands, as in the following sentence translated into Bulgarian: *Big companies in Bulgaria use Google's services – Големите компании в България използват услугите на Google.*

The difference between *Translate* and *Transliterate/Transcribe* may be important for clearer instructions on proper nouns or products consisting of words that also have a lexical meaning. It is common for proper nouns to be transcribed phonetically from languages that use Latin script to the languages that use Cyrillic script such as Bulgarian. Therefore, the simple tag *NO* in the *Translate* column might confuse the translator as to whether a particular unit should be left in Latin script or not.

The results showed that 84 terms could be transliterated or transcribed phonetically and that this method could be used to guide glossary translators more accurately. Moreover, it occurred mainly with the names of persons (e.g. some characters in the games). However, it may be necessary to add this column and to fill it in separately for each target language, as there is a transliterated/transcribed version for some brands in certain countries and this needs to be checked by the translator of the glossary.

2.4 Conclusion

The way in which the internship was structured, as well as the involvement in the company's culture, especially the participation in the company's general meetings, played a crucial role in promoting engagement and improving performance with increased enthusiasm. The collaborative atmosphere fostered by the opportunity to share the workplace with the other interns and the employees of the company was important for all interns to develop new ideas and critical thinking.

The internship provided a dynamic and varied experience where the interns had the opportunity to familiarise themselves with the company's tools, gain insights into the functioning of the MT engine and workflow, analyse linguistic resources and make recommendations for their improvement. The highlight of this experience was the interaction with the Product team, who not only appreciated the ideas suggested, but also integrated some of them into subsequent updates of the tools and resources. These ideas were also fully documented in the reports produced for each assigned task.

The comprehensive understanding of all tools and resources in the first part of the internship proved to be extremely important for the successful execution of each individual task during the second part.

In my case, the second part focused on the task of curating a glossary of a client. In the process, I learned about the methods used to create glossaries and the criteria for including or excluding certain words. This increased my interest in finding patterns in the glossary, which led to a quantitative analysis of the PoS categories of glossary units and the origin of these units. Examination of the glossary led to reflections on the singular and plural forms of the glossary units and on the need for additional information in the case of different scripts of the SL and the TL (e.g. Latin and Cyrillic).

Considering that the glossary system is used by both human translators and a MT engine, particular challenges arise for the successful selection and translation of terms, the functioning of the system and its future development. This topic proved interesting and valuable for further analysis, especially the translation of multiword units that include some creative effects and require special attention. To get a complete picture, we need to understand not only the historical context and the state of the art of MT, but also the differences between glossaries used by humans and those used by machines. This information can be found in the next chapter.

3 Theoretical Overview

This chapter introduces the main theoretical concepts related to machine translation, Computer-Aided Translation (CAT), the formation and translation of creative elements in multiword terms, and the creation, curation and translation of multilingual glossaries.

A brief introduction to the history and current state-of-the-art of MT systems and the relevant CAT tools currently used in the translation industry is given in Section 3.1. Some approaches to terminology management and some requirements for glossary creation are presented in Section 3.3. Section 3.4 deals with relevant aspects related to the understanding and translation of Multiword Expressions for Special Purposes (MWESP). In this chapter we will also identify some issues related to the translation of creativity (see section 3.4.2), an aspect relevant to the translation of the units we have chosen for a more detailed analysis in this report.

The above topics form the theoretical basis of this internship report and need to be addressed in order to provide a context for and deep understanding of the work presented in Chapter 4 and Chapter 5.

3.1 Machine Translation and Computer-Aided Translation

The existence of Machine Translation (MT) and Computer-Aided Translation (CAT) tools has led to a radical change in the translation market over the last three decades as far as the execution of translations is concerned. The translation market is becoming increasingly demanding in terms of the speed of the translation process, which in many cases requires the use of CAT tools, MT or a the combination of both. Some translators, especially in the area of technical translation, are usually required to use CAT tools. The growing demand for new approaches and research in the field of machine translation has led many research groups to turn their attention to it, and the results are promising. However, the goal of MT (to create a system, which produces a fully automatic high-quality translation) has not yet been reached.

With regard to the translation workflow at Unbabel and the software tools involved, a brief summary of the history and functioning of MT systems and CAT tools, as well as a brief overview

of the software currently available on the market, is necessary to place the technology used at the company in the context of contemporary systems: Unbabel uses neural machine translation and has several internal CAT tools that support human post-editing of MT output.

3.2 Brief history and functioning of machine translation

In this section, we follow Daniel Stein in a brief overview of the history, typology and functioning of MT systems, as presented in his article "Machine translation: Past, present and future" (Stein, 2018).

According to Stein (2018), in practice, the history of MT began with the appearance of the first computers after the World War II, although the philosophy of machine translation and the idea of representing a language in a formal way emerged in the 13th century with the ideas of the Catalan philosopher Ramon Llull (1243 – ca. 1316). Llull had a theory in which the reasoning of God and the world were objectified by means of a formal language. Later, Gottfried Wilhelm Leibniz (1646 – 1716), a German philosopher and mathematician, used this idea to develop a theory which attempted to define a set of the smallest units of meaning in order to compose all thinkable thoughts. Stein (2018) categorises these ideas as a philosophical school that aims to create a universal language, and he distinguishes it from schools of thought, directed to create secret languages and codes, such as that of the German physician and alchemist Johann Joachim Becher (1635 – 1682). Becher developed a system in 1661 which represented the idea of inventing a new language that could enable speakers of all languages to understand any other language after a special one-day instruction. This approach was similar to the first technical approaches in MT in the late 1940s and was based on dictionaries whose units were labelled with numerical codes and thus related to units in other language dictionaries. In the context of the World War II and the Cold War, cryptology benefited from these ideas, as communication in secrecy was much more important at that time than creating a universal language for communication without borders.

Formally, according to Stein (2018, p.6), the "birth of MT", can be seen in an exchange of letters between Warren Weaver and Andrew Booth following the decipherment of the German ENIGMA, which was undertaken using statistical methods and computing machines. In these letters, Weaver introduced the idea that "a book written in Chinese is simply a book written in English which was coded into the 'Chinese Code'" (Weaver, 1955, as cited in Stein, 2018, p. 6) and suggested that the methods used by cryptography could be useful for translation as well.

Nevertheless, that appeared not to be true, as these methods were inadequate for more complex translations that required syntactic changes and not just a direct replacement of lexical units. Undoubtedly, the translation process differs considerably from the deciphering of a code, as Hutchins (1986) states: "the decipherment of the highly complex Enigma code [...] was not translation; it was only after the German texts had been deciphered that they were translated". Cryptography, on the one hand, consists of encryption and decryption, which are "the deterministic processes of substituting, ordering, and permuting discrete inscriptions" (DuPont, 2018). Translation, on the other hand, is not based on the direct substitution of characters or words, but on the transfer of units of meaning from a SL to a TL, which are usually not limited to single words. Thus, in cryptography, a unit of the original message, for example a character, is replaced by another in the encrypted message, whereas in translation, a semantic unit, such as a word in a SL, may correspond to one or more words in the TL and vice versa. The latter also involves complex transformations related to word order, grammar and punctuation rules of the TL and so on.

The first systems for MT were based on dictionaries and some simple syntactic operations, i.e. they applied some simple rules of the target language instead of copying the structure of the source text, for example. However, this approach essentially served to prove that machine translation was possible. As a result, a lot of money was invested in research and development of MT systems until 1966. However, following a famous report by the Automatic Language Processing Advisory Committee (ALPAC), published in 1966, which stated that no significant progress had been made in MT research and that it would therefore not be of use in the near future, funding was immediately and drastically cut and did not resume until the 1980s. In the same report, ALPAC recommended increased spending on the development of tools that would support, rather than replace, the translator. This led to the development of CAT tools, which are explained in more detail in section 3.2.2.

The MT systems developed in the period between the 1960s and 1980s were based on linguistic rules and are now known as **Rule-Based Machine Translation (RBMT)** systems. As Stein suggests, there are three types of RBMT: **direct** (based on a direct replacement of words using parallel dictionaries), by **transfer** (the most common type of RBMT, based on a set of morphological, syntactic and semantic rules, combinations and exceptions elaborated by teams of linguists) and by **Interlingua** (based on a *neutral language* - a metalanguage through which the meaning of both the source and the target language could be expressed). However, the creation of these systems required significant investment and time, as they had to be created manually by experts in both source and target languages. Even after years of developing RBMT

systems, the results delivered were not good enough, due to contradicting rules, internal conflicts or the coverage of the system.

In the 1980s, funding and research in MT were revived following a promising article by Peter F. Brown (Brown et al. 1988, as cited in Stein, 2018), which changed the point of view and suggested a return to statistical methods. By this time, a large amount of machine-readable parallel corpora were available, allowing the creation of **Statistical Machine Translation (SMT)** systems. These systems worked by calculating and comparing the probabilities of co-occurrence of words or groups of words in aligned bilingual corpora. They were quick to build – with the necessary corpora, they could be created within a few days – and required no knowledge of the source or target language. The two types of SMT, **word-based** and **phrase-based** SMT systems, operated by comparing words or sequences of words. Yet, the first ones could not identify multiword expressions and the second ones, although able to do it, could only identify sequences of consecutive words. The best results were reported when translating in specialised domains.

Some of the problems with SMT systems were that they produced ungrammatical sentences in many cases, required laborious manual correction, and the results provided were not good enough when dealing with language pairs with very different language structures. Moreover, they reached a limit in terms of performance, as the problems could not be solved even with larger corpora. As a result, hybrid systems appeared which combined the advantages of SMT with some of the rules of RBMT systems (syntactic preprocessing, semantic disambiguation, etc.). These systems were the state-of-the-art for 30 years. However, despite intensive research in the area of SMT systems, the results were not perfect and improvement was slow and not very efficient.

3.2.1 Neural Machine Translation

In recent years, another important method has emerged in computer science, which was based on research papers dating from the 1940s and 1950s regarding *artificial neural networks*. Artificial neural networks work as a framework for data processing and are used in various machine learning algorithms. Neural networks are inspired by biological neural networks in human and animal brains, which consist of neurons, links between neurons, and their synapses.

In general, computer systems are procedural, i.e. they work in a linear way: after executing one line of code, the programme moves on to the next and so on. According to Shiffman (2012), this is the most salient difference between the linear model and artificial neural networks, in

which the "information is processed collectively, in parallel throughout a network of nodes" (2012, p.446). These networks also have the ability to adapt themselves (to "learn") by changing their internal structure in relation to the information being processed. This means that they can make adjustments to their internal structure over time. Nowadays, these systems are increasingly used for various applications, such as classification, recognition and identification of data and much more.

According to Koehn (2017, p.5), the idea of incorporating artificial neural networks into machine translation systems emerged during the first wave of research in neural networks in the 1980s - 1990s, but the complexity of such systems exceeded the possibilities of the computational resources available at the time, and therefore the idea was not implemented for almost two decades.

The renewal of research in this area began in the 2000s with the incorporation of neural language models into traditional statistical machine translation systems. These methods required specialised hardware and experience to be developed and tested, resources that many research groups did not have. In the 2010s, however, neural network methods began to be used in a variety of other components of SMT systems in addition to their use in language models, such as, according to Koehn (2017), "providing additional scores or extending translation tables [...], reordering [...] and pre-ordering models [...]". Also, neural networks methods have been used in joint translation and language model development by Devlin et al. (2014).

First steps in pure neural machine translation systems (i.e. not based on an SMT system) were made and they provided relatively good translations for short sentences. Only after various research groups added some components, such as *attention mechanisms* (which selectively focus on certain parts of the source sentence), did neural machine translation systems surpass traditional SMT systems and become the new state-of-the-art.

Some of the toolkits currently available for research and development of neural machine translation are *Nematus, Marian, OpenNMT, xnmt, Sockeye, T2T*, etc.

There are also ready-made neural machine translation engines that can be used for free online. The most popular at the moment are *Google Translate* and *DeepL*, which include various language pairs.

There have been attempts to integrate glossaries into NMT engines. Hasler et al. (2018) propose a method for including terminology constraints into NMT, which involves decoder attentions. The authors report that results on four language pairs show that terminology constraints can be respected during NMT decoding while maintaining the overall quality of the translation.

3.2.2 Computer-Aided Translation Tools: functioning and use

As mentioned on p. 43, the 1966 ALPAC report was important for the evolution of MT and it focused on the impossibility at that time of significantly improving the quality of MT output: "As it becomes increasingly evident that fully automatic high-quality machine translation was not going to be realised for a long time, interest began to be shown in machine-aided translation" (ALPAC, 1966, p.25). Early researchers in the area, such as Bar-Hillel, suggested the same. According to this author, there were three possible directions for further investigation: "(1) Machine-aided human translation, (2) man-aided machine translation, (3) low-quality autonomous machine translation", (Bar-Hillel, 1971, p.75).

All three directions have been further explored, for different purposes. Nowadays, **autonomous machine translation** is used, for example, to understand the general idea of a message, as Bowker and Ciro (2019, p.23) point out:

> [...] raw machine translation may be used for "gisting," which means that a user can employ machine translation for personal use in order to get the gist or comprehend the general idea of the meaning of a text that has been written in another language.

In many international companies where a quick translation is needed, autonomous MT can be used. These cases usually involve written communication between employees. In such cases of internal communication, free online services, such as *Google Translate* and *DeepL* are often used (see also page 45). MT is not normally used for publications or communication with customers.

Man-aided machine translation is another solution for improving the quality of MT. It can include one or more of the following approaches: pre-editing MT input; controlled language; customisation of the machine translation system or process; or post-editing MT output. Controlled language in MT refers to the use of a specific restrictive grammar and vocabulary when creating or editing a text with the purpose of avoiding complex structures, ambiguity, unclear meaning and thus incorrect translations. As this approach is highly demanding in terms of creation, implementation and training of specialists, according to Bowker and Ciro (2019, pp. 60-61), other approaches, namely pre-editing MT input, can also lead to more accurate MT output by "simply trying to remove ambiguities and constructions that could pose difficulties for a machine translation".

Customisation refers to the addition of manually customised tools, such as glossaries, which are applied to the machine translation and affect the final output. Customisation may determine, for example, the tone (formal/informal) of the final output or the selection of a preferred

equivalent of specific units. For example, when glossary units are added (one source unit must correspond to one target unit), the preferred target unit is used to automatically substitute another possible translation of the same source unit. The final result will automatically appear in the machine translation output.

The translation engine of *DeepL*, for example, includes such a customisation feature for certain language pairs. This feature includes, among other things, the possibility to add user-made glossaries (see section 3.3.2) and to select the tone of the translation (formal or informal) (DeepL, 2022b).

The company hosting the internship related to this report, Unbabel, also offers customisation in a similar way, especially with regard to the glossary system, which is the focus of this work, as described in section 2.2.2.

Post-editing involves correction of the errors that appear in MT output. We are going to focus on this process as it is also one of the approaches used at Unbabel. The description is presented in section 3.2.3.

Machine-aided human translation, also called **computer-aided translation** (**CAT**) comprehends different types of software developed to assist professionals in the translation industry in the manual execution of translation tasks.

This includes all kinds of tools specifically designed to facilitate the translation process. One of the most popular tools are translator's workstations/workbenches (also called translation environments), which generally incorporate a resource called Translation Memory (TM) and additional tools that assist human translation. They can be used in combination with separate MT systems and some of them even include a MT component.

Considering the fact that in MT systems the translator cannot intervene in the actual translation process (although pre- or post-editing of the source/target text can be performed by a human agent), CAT tools are designed precisely to provide translators with user-friendly software to help them perform a translation task manually. This is achieved with the aid of various components, "translation memory systems or similar tools, but usually also [...] bitext aligners, terminology management systems, term extractors, active terminology recognition tools and bilingual and/or multilingual concordancing functions" (Collection of Electronic Resources in Translation Technologies, 2012).

The difference between MT and CAT is summarised by Bowker (2002, p.4):

> The major distinction between MT and CAT lies with who is primarily responsible for

> the actual task of translation. In MT, the computer translates the text, though the machine output may later be edited by a human translator. In CAT, human translators are responsible for doing the translation, but they may make use of a variety of computerized tools to help them complete this task and increase their productivity. Therefore, whereas MT systems try to replace translators, CAT tools support translators by helping them work more efficiently.

According to Reinke (2018, p.56), the first usage of CAT tools occurred in the 1960s when the "European Coal and Steel Community developed and used a computer system to retrieve terms and their contexts from stored human translations", a task accomplished by identifying matching lexical items in source and target sentences. This system was more of a terminology management system because its main feature was the retrieval of terms and their context. It differs from today's CAT tools, which, like TM systems, focus on reusing existing segments of human translations to facilitate and speed up the translation and revision process. With some of them, even MT outputs are available automatically.

Reinke (2018, p.60) points out that "the core functionalities of commercial TM systems have remained very much the same since the first – mostly still MS-DOS-based – applications became available at the beginning of the 1990s". However, a major change in these systems occurred in "the way the translation processes are organised and the way the parties involved in these processes interact and collaborate" (*Ibid.*, p.61), namely with the inclusion of the possibility of real-time online collaboration. This allows different members of the team responsible for a translation project to work on it simultaneously, reducing the time needed to create, edit, approve and deliver the finished translation.

The way TM systems work is simple. Before the translation process takes place, the source text is automatically divided into segments and empty cells are provided for the translator to fill in with the corresponding target text segments. During the translation process, these aligned source and target segments are stored in a database called a translation memory. This database contains aligned source and target segments and the TM system uses it to perform an automatic search for similarities in the source segments. These can then be found again in the same text or in a subsequent translation project. If a similarity above a certain threshold is found (i.e. a "fuzzy match"), the system fills in the target text cell with the stored one so that it can be modified manually by the translator. In case of identical source text segments (i.e. an "exact match"), the tool automatically uses the saved segment. The translator has the option of reviewing and accepting the target segments or modifying them.

Therefore, the more segments are stored in the database, the higher the probability that the TM system will find source segments in new texts that are analogous to the ones stored in the

database. In this context, it is important to note that translation memories should be organised per domain, client and/or other type of category. In this way, the search can only be performed within a chosen category, which reduces the probability of false positive results.

Today, there are different types of CAT tools on the market, such as the workstations (like the workstations *SDL Trados Studio*, *memoQ*, *Wordfast*, *Déjà Vu*, *Across*, *OmegaT*, *MateCat*), as well as tools that perform automatic linguistic quality assurance or tools used for terminology management. They are either incorporated into a translation memory tool or workstation, or exist as a stand-alone software. Some of them store the files in a local directory and do not depend on an internet connection. Others are cloud-based and facilitate online collaboration between different members of the translation team and project managers. The translation memory system used at Unbabel is mentioned in section 2.3.1.1. It works almost like the TMs included in the translation workstations available on the market: they identify identical segments of the source text in the TM and apply an existing target segment. An important difference is that with translation workstations, the inclusion of the segment in the TM usually happens automatically after the segment has been approved by the translator. At Unbabel, however, there is a separate step of TM curation. In this step, professional translators review the selected segments in an Excel spreadsheet. They decide whether a specific segment is suitable for storage in the TM database or whether a correction is required. If a correction is necessary, the professional translator applies it. The TM system is then "fed" with the approved source and target segments. As far as retrieving data is concerned, the TM systems at Unbabel, as already mentioned (see section 2.3.1.1), only work with exact matches. This is usually not the case with TM systems on the market. These TM systems can detect a similarity between the segments in the database and the source text, even if no exact match is found, and assign a percentage value of the similarity (i.e. a fuzzy match).

3.2.3 Post-editing of Machine Translation

According to Bar-Hillel (1960, p.93):

> [...]full automation of the translation process is incompatible with high quality. There are two possible directions in which a compromise could be struck; one could sacrifice quality or one could reduce the self-sufficiency of the machine output.

This statement shows that even the early researchers in the area had the idea that the evolution of MT would not be instantaneous and that the high quality of translations produced

by professional translators could not be achieved by machine translation systems without the interference of a human editor.

This is how the idea of the post-editing process came about, as mentioned by ALPAC (1966, p.65): "A system such as that [...] might properly be called human-aided machine translation, since the post-editing process was added after it became apparent that raw output was unsatisfactory and since humans are employed essentially to make up for the deficiencies of the computer output."

The basis of the Unbabel workflow is a combination of MT systems, CAT tools and post-editing of MT output. In Figure 2.1 (which is replicated below as Figure 3.1 for ease of illustration) we can see where exactly these processes are included in the case of Unbabel: As mentioned

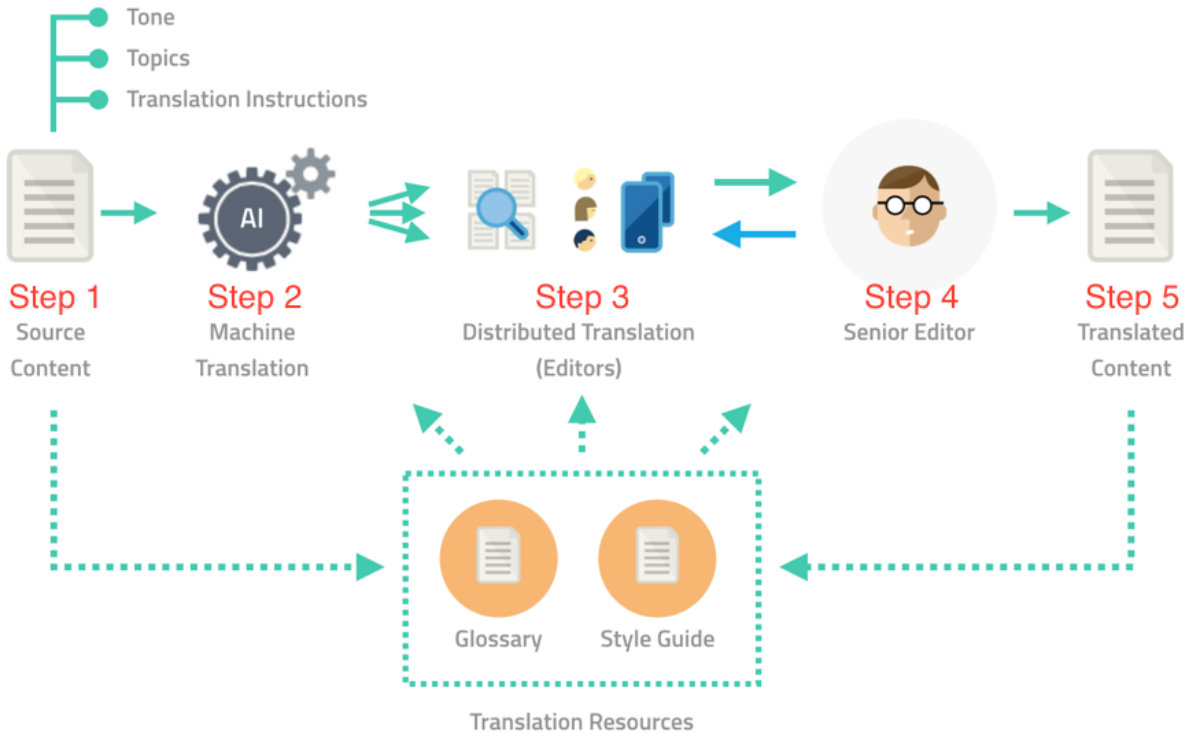

Figure 3.1: Replication of Fig. 2.1. "A model of the Unbabel translation workflow"

in section 2.2.1, "Step 1" includes formatting and checking the source text, namely, the following automatic processes: deletion of unnecessary white spaces, anonymisation of sensitive and/or personal data, search for matches in the glossary and in the TM databases. "Step 2" is reserved for machine translation and some additional tools, namely the software that chooses which MT variant should be used (the selection is between grammatically correct and ungrammatical variants) and the quality estimation system that assigns a score to the MT output. "Step 3" involves human post-editing of the MT output, client's instructions (on the required tone, area of busi-

ness and other requirements), as well as some additional CAT tools, which highlight identified glossary terms (in both source and target text) and any potential issues in the target text that might require special attention (*Smartcheck* tool). Unbabel's *Smartcheck* tool is presented on page 18. As mentioned there, it is a system that contains a set of rules of the target language grammar and is applied to the target text in order to identify any possible issues. This system, which is used in the context of post-editing of machine translation, only highlights potential issues, but the final decision as to whether something in the text should be corrected or not rests with the post-editor.

As Bowker and Ciro (2019, p. 38-39) point out: "Gradually, machine translation and translation memory systems have become integrated, allowing professional translators to make use of both types of technology." Translators and translation companies nowadays have the option to combine different tools in the name of a greater productivity rate or better quality of the translations and, why not, a certain level of alleviation in some particularly challenging translations. As explained here, the combination of MT systems with CAT tools mentioned at the beginning of this section is currently widely used, which is also the case of Unbabel.

3.3 Terminology Management and Glossaries

In this section we will briefly discuss terminology. Its importance lies in the types of glossaries we are considering. They are multilingual glossaries used in specialised fields and domains where translation plays a central role. This leads us to the need to observe how terminology has resolved these topics in the past and to observe possible translation strategies. With regard to translation and terminology, Cabré (2010) clarifies the relations between these two areas and how they depend on each other. Cabré (2010) analyses the terminology field from a linguistic, cognitive and communicative perspective and mentions the different approaches for solving terminological challenges in translation.

Thus, for Cabré (*op. cit.*), from a linguistic point of view, terms are lexical units that introduce specialised meanings when used in specific contexts. From a cognitive point of view, the terms are conceptual units that are a representation of specialised knowledge and form a structured cognitive view of a certain specialised field. In a communicative perspective, this author argues terms are used to share expertise between participants in the field or with the general public.

This author points out that the field of terminology studies the terms, deals with their

collection, analysis and standardisation and organises them in glossaries and databases (2010, p. 357):

> Terminology aims at collecting specialized terms to compile them and produce terminological resources (glossaries, dictionaries, vocabularies or databases) intended to be readily accessible and useful to translation experts, among other professionals.

From the point of view of translation, Cabré (*ibid.*) defines terminology as an *instrument* that provides terminological resources for the translation process, such as information about the terms, equivalents or different translation possibilities. It also helps the translator to acquire specialised knowledge and organise it.

Looking deeper into the terminology field and trying to differentiate the concepts, we can mention Rogers (2006) who states that *terminology* can be used in four different meanings. It can refer to the concept-based methods used to compile terminologies and, in particular, to the principles by which this is done. It can also refer to the application of terminological methods (known as *terminography*); to a collection of terms in a particular field; and to the published (electronic or printed) version of that collection.

According to Rogers (2006), "[...] a terminology can be paper or electronic, as also 'glossary,' for example". However, the author points out that there are two types of terminology resources that are used and stored exclusively on an electronic medium, namely *termbanks* (also known as *terminological data banks* or *terminology data banks*) and *termbases* (i.e. *terminological databases* or *terminology databases*).

It appears that the differences between *termbanks* and *termbases* are somehow blurred.

For example, in his article, Rogers (2006) mentions that some of the main differences between termbanks and termbases include the following: termbanks started to appear in the 1960s and 1970s and are normally more complex and built by large entities, such as non-commercial organisations (or are a result of interinstitutional collaboration), while termbases are more recent and are the product of the availability of commercially distributed tools, such as *terminology management systems*. Termbanks are intended to be accessed and used by a vast number and type of users such as translators, engineers, students, to name a few, while termbases are used by freelance translators or in-house translators and it might not be possible to share them depending on the file format used.

One of the examples of termbanks mentioned by Rogers is IATE (Interactive Terminology for

Europe, also known as Inter-Agency Terminology Exchange) [1]. However, on its official website it is categorised as a terminological database.

Melby (2012) distinguishes termbases and termbanks based on their size:

> There are many types of termbases in use, ranging from huge termbases (usually called 'termbanks') operated by governments, to medium-size termbases maintained by corporations and NGOs, to smaller termbases maintained by translation service providers and individual translators.

This shows that the definition and the use of those two terms is not consistent throughout the field.

According to Cabré (2010:362), "The primary advantage of terminology banks in relation to traditional glossaries is the possibility to be continuously updated, as well as their capacity to store a large number of terms and term-related information, which allows oriented and selected data retrieval".

We consider glossaries to be lists of words that contain information about the concepts they represent and/or their equivalents in foreign languages. These words may be part of the general vocabulary or used in specialised contexts. Often, their description in the specific context is also included, which shows the main objectives of glossaries: to define the selected words, to provide translations (if this is the case), to disambiguate and to explain the word in the required context. Glossaries can be understood as context-specific lists of selected words and are not intended to provide thorough knowledge of all possible uses and contexts, which distinguishes them from dictionaries, for example.

Glossaries can also be created and used by students learning foreign languages, for example, so that their knowledge and vocabulary about a certain topic or area is formed and consolidated. They can be used by translators or professionals of other areas to organise the terminology used in a certain text, domain or by a certain client. For large-scale technical and scientific translations, a glossary is used to ensure that the same translations are used throughout the texts. However, when the number of the glossary units becomes higher, it is no longer practical to create, use and manage it on paper or by working in a text file.

For this reason, many professionals, such as translators, interpreters, technical writers and others, use digital tools to manage terminology - Terminology Management System (TMS).

[1] Available at: https://iate.europa.eu

Currently, there are many such systems. They are useful for collecting, maintaining, retrieving, archiving and sharing terminological data. Some of them, such as *SDL MultiTerm*, can be used separately or incorporated in translation environments. Others, such as *InterpretBank*, are designed in a way to be used by interpreters during their assignment. Currently, many TMSs have the option of storing the information on the internet (cloud service). This makes sharing and accessing the terminological database easier, both for different members of the translation team or between freelance professionals and may lead to more secure storage of data.

Many of the available TMSs allow terminology databases to be exported in XML format, which is the preferred file format in the industry. The benefits, such as facilitated exchange between translators and companies, are described by Roturier (2019).

The fundamental characteristics of terminology management are pointed out by Sprung and Jaroniec, (2000, p.XVIII):

> Terminology management typically involves a system for cataloguing, updating, retrieving, and managing terms. Without it, there would be far greater danger of using incorrect or inconsistent terminology. Terminology management is critical for reducing cost and turnaround time on translation projects: if translators are given standard, approved glossaries of terms on which to base their translations, quality will be far higher with greater consistency.

While they mention the importance it brings to translation performed by humans (see section 3.3.1 for more details), there are also advantages for machine translation and computer-aided translation, which are explained in the following sections.

3.3.1 Glossaries for human use

In the following paragraphs we will examine the creation of glossaries used by humans, through a case study, namely the project of the company "Ericsson". In doing so, we will see a possible model for creating a glossary, the required elements within the glossary and some solutions to problematic points.

Jaekel (2000) describes a terminology management project for commercial purposes in the field of telecommunications. The project, developed in the 1980s and 1990s, involved translators and other professionals, none of whom were trained or experienced terminologists. This led them to search for and experiment with different approaches.

Naturally, there were different stages in the identification, selection of terms and the compilation of the terminology database. The first stages were carried out by the translators who worked

for the company. They identified and extracted candidate terms from the translations they had done. Then a translator had to submit a candidate term into the terminology management software, and then this term was reviewed by a professional with a higher access level.

There were three classes of term acceptance, which were: “Class 1”, “Class 2” and “Class 3”. In case a term candidate (“Class 1”) is accepted, it is promoted by the second professional to “Class 2” – “To be verified” and is later reviewed by an admin who has the highest access level. This professional has access to promote the term to “Class 3” – “Fixed” and to include it to the reviewed and accepted terms in the database.

This procedure was to ensure that the accepted terms are reviewed by several professionals, who checked whether they were relevant to the database and whether the information given about them, such as the domain and description, was correct.

The use and importance of domains is one of the main characteristics of the terminology management used for translation made by humans (including computer-aided translation). For example, as Jaekel (2000, p. 159) mentions, one of the early versions of the "Eriterm" glossary included the Swedish source terms, the target terms in three languages and the domains: "Each of some 13,000 term records contained a domain, but no definitions". This shows significance of including the domains, especially for glossaries with larger scope.

Another important point of this experience is the way the domains and subdomains were organised. Jaekel (2000, p. 165) reports that the initial abundance of considered domains led to the need to create more general domains: these included the previously existing domains as subdomains. This was necessary because managing so many domains of the same level became too difficult. Grouping them made it easier to search, use and manage them: for example, the subdomains "chemistry", "geography" and "linguistics" were organised in a higher-level domain called "Sciences" (Jaekel, 2000, p. 166).

Other characteristics of the glossary can be found in the “Term record structure”, which Jaekel describes. This term record includes technical data, entry class (1, 2 or 3, as mentioned above); domain and subdomain; definition and information about the definition such as a source; and others. The mentioned refer to the entire term record, while there is also a language specific term record, which can include information about the usage of the target term (language variant for example); some grammatical information; source of the target term, acceptability; customer name; and others.

The glossaries that are the focus of this work are glossaries specifically created for translation purposes, such as Jaekel’s case study mentioned above. However, there are other types

of glossaries for human use, such as monolingual glossaries created by experts in a particular field. They differ from the glossaries considered here not only in the absence of an equivalent term in another language, but also in the starting point for terminology work, as Cabré (2010, p. 357) explains: "For specialists, the concept is the starting point for terminology work, while for translators the concept is the intermediate point between the original term and its equivalent".

While it is true that translators working with specialised texts are in many cases also experts in these fields, this is not always the case. This is important to understand because in a monolingual glossary of a specialised field, the main unit around which glossaries are built is a concept, whereas in glossaries created for the specific purpose of translation, the main unit is the source term, which must find its equivalent in another language. In some special cases, the source unit has no equivalent in the target language. This leads us to the idea expressed by Cabré (2010) that a translator can use two types of strategies to solve terminological challenges in the texts to be translated. He can participate in the terminology work as an observer or as a creator.

If a translator takes on the role of observer, in cases where an equivalent term exists in the target language, a digital search can be carried out in the numerous terminology databases that exist on the Internet, such as IATE (Interactive Terminology for Europe) or the Microsoft Terminology Collection or other private databases. The role of creator should only be taken on if absolutely no equivalent can be found in the target language. After all, if a translator coins a new term that is not yet known among experts in the specialised field, there are two possible problems: The meaning of the term and the text could either be lost or changed; the term could never become known and used among the experts in the relevant field. An important note that should be mentioned is the amount of work that is done today without the help of digital tools for glossary creation and translation. In the past, the entire procedure of collecting lists of terms, translating them was done manually. However, as mentioned earlier in this section, in Jaekel's case, digital tools were already available and used for the project, which started in the 1980s.

Cabré (1999, p. 163) also touches on this topic in the end of the 1990s:

> "Computer-aided text analysis and the possibility of processing large amounts of information have changed the bases of terminology compilation, as well as how the appropriateness of terms is conceived, and the degree of human intervention in the whole work process."

Nowadays, the use of digital tools, such as terminology management systems or even very simple word lists in a digital environment, such as Excel spreadsheets, is almost obligatory in one form

or another. This also applies to the entire translation process, so glossaries used in the translation process are no exception.

As explained in section 2.2.2, Unbabel uses various digital tools to create and translate glossaries. One of these is an online glossary platform that was developed in the company specifically for this purpose and is directly linked to the platform on which the texts are translated. Since these glossaries are used by both humans and machines, in the next section we will go into a bit more detail about the type of glossaries used by machines and analyse some examples.

3.3.2 Glossaries for machine use

According to Bowker and Ciro (2019, p.46): " [...] regardless of the approach employed, machine translation systems continue to grapple with the fact that language is inherently ambiguous." In this regard, one possible way to deal with ambiguity in the use of glossaries during or after MT. Thus, a specific target unit can be previously selected and linked to an ambiguous source unit, depending on context.

The creation and application of glossaries is different for human use, machine use or mixed use. The challenge of having machines disambiguating linguistic units is mentioned by Bowker and Ciro (ibid.):

> "It is important to keep in mind that a word or a sentence can be ambiguous for a computer even if it does not seem to be ambiguous for a human. People can draw on real-world knowledge to interpret meaning but RBMT systems cannot. They can only follow the linguistic rules."

The main difference in using glossaries created for human translators and for machine translation engines is the information necessary to take a decision on which term to use. By this I mean that glossaries created for humans contain additional information that stimulates and facilitates the decision-making process.

In contrast, glossaries created for machine use do not require additional information and simply copy the target term from the database and apply it in the target text.

In them, the decision-making process carried out by humans is moved to the time before the source text is translated, rather than during the translation. This process includes the creation, curation, translation and review of the glossary that will later be used by the MT system. The curation of glossaries at Unbabel is explained in detail in section 2.3.5. This is an important step because although these glossaries are intended for "mixed" use (by both humans and machines),

they do not provide any additional information to the post-editor working on an automatically translated target text that needs a review. Only the source and target terms appear (the target term is already applied in the text). As mentioned earlier, the curation of the glossary should therefore be done carefully and in a way that anticipates any possible issues or ambiguities.

The process of creating a glossary for machine use involves automatic extraction and/or manual selection of the units to be included in the glossary. This first step is done using existing texts on the selected topic or area in the source language (in the case of Unbabel, this is English). These texts may also include company documents, which are necessary to capture the exact terminology used by the company. In some cases, marketing teams provide strategies to understand and define the company language and terminology, as well as the "voice" of the company. This process can relate to the promotion of the company or a particular brand and (in some cases) influences sales.

This extraction of candidate terms can be done by automatic statistical systems that include certain rules or tags so that, for example, only nouns or noun phrases are included in the results.

Some glossaries for machine use, e.g. those used at Unbabel, recognise an exact match in the source text and apply the target term unchanged. For this reason, in those type of glossary systems, multiple forms of the same term must appear and verbs are not usually accepted because of the large number of possible inflexions. However, in specialised fields, verbs and verb phrases can also be included in glossaries.

In the case of Unbabel, singular and plural forms of the noun must appear ("apple", "apples") in a glossary created with English as the source language. If only one of the two forms is included in the glossary, only that form will be recognised by the system, as these glossaries work with exact matches. At Unbabel, for example, all glossaries are created in English (English terms are the source terms) and then translated into other languages (target terms). As far as we were informed, the glossary system usually uses the English terms as source terms. However, it can also use the glossary target terms in another language of the glossary as if they were source terms. Thus, for example, it will consider the Bulgarian target terms of the glossary to search for exact matches in a Bulgarian source text and provide the English correspondence of the term found in the glossary. As already mentioned, the system can only identify exact matches.

A challenge which may arise in this case is that Bulgarian (like many other languages Unbabel works with) is morphologically very rich. In Bulgarian there are no grammatical cases, but the definite article is attached to the noun or the adjective as a suffix. For example, if for a given noun in English only two other entries are necessary (singular and plural), in Bulgarian there

can be 4 or more. Take for example the word "Portuguese" only as a noun and in its meaning of "a person from Portugal" [2]. In Bulgarian, for example, there are 4 feminine forms for this noun:

"португалка" (a Portuguese woman)

"португалката" (the Portuguese woman)

"португалки" (Portuguese women)

"португалките" (the Portuguese women).

The number of masculine forms is even greater, since the definite article can have a full and a short form, depending on the role of the noun in the sentence (similar to grammatical case in other languages):

"португалец" (a Portuguese man)

"португалеца" (the Portuguese man, short form of the definite article)

"португалецът" (the Portuguese man, full form of the definite article)

"португалци" (Portuguese men or a group of Portuguese men and women)

"португалците" (the Portuguese men or a group of Portuguese men and women).

There is also a neutral form that can be used for some nouns or when you are talking about small things, children, or expressing affection:

"португалче" (a Portuguese child)

"португалчето" (the Portuguese child)

"португалчета" (Portuguese children)

"португалчетата" (the Portuguese children).

Thus, in the case of an exact-match glossary system, 13 Bulgarian forms can be included, while in English there are only two. This example ignores other meanings of "Portuguese" such as the language, for example, and also the fact that the adjective has the same spelling, which can lead to problems, described later in the text. For languages with grammatical cases, the number of forms needed in the glossary would be even higher.

On the other hand, this very noun should probably never be included in a glossary for use by a machine translation engine, as it is not a term. However, many terms, which would need to be included in such a glossary do have multiple forms.

[2] As it appears here: http://dictionary.cambridge.org/dictionary/english/portuguese (Consulted on 26.01.2023)

At Unbabel, for example, the glossaries have English source terms and are then translated into other languages. The same glossaries can also be used in texts that have Bulgarian as their source language. Therefore, the Bulgarian target terms are assumed to be the source terms and the system would look for exact matches between the database and the source text. For the reasons explained above and because of the multiple inflectional possibilities in some languages, such as Bulgarian, exact matches will be rare. In some cases, as in the example above, the probability of identifying a Bulgarian term in a source text is 1 in 13. This means that when a glossary is used in reverse (the Bulgarian target terms are assumed to be the source terms), exact matches tend to be the exception and therefore the usefulness of a glossary used in reverse is doubtful.

A system based on exact matches is not appropriate for verbs, adjectives, verb phrases, etc. for the same reasons. The combinations and morphological differences will be too great, making the glossary inefficient.

The situation is different with glossaries for machine translation, which automatically inflect the word and can be used for nouns, verbs and others. An example is the glossary system of DeepL, which currently is only available for a few language pairs.

According to the documentation on the website (DeepL, 2022a), DeepL's glossary can be used both for words and phrases, without specifying which category of words it is. However, after a short test, it becomes clear that nouns and verbs are correctly inflected for the English-Italian language pair and according to their number or number and tense.

Consider the following example, tested on December 15, 2022. After defining the verb "love" as a glossary unit and choosing the verb "amare" as the target equivalent, the following results appear:

"I love you" – "Ti amo" / "Io ti amo"

"We love you" – "Ti amiamo" / "Vi amiamo"/ "Noi ti amiamo"

"We loved you" – "Ti abbiamo amato" / "Vi abbiamo amato" / "Noi ti abbiamo amato".

On the platform it is clear that the system has recognised the glossary unit because in this case the recognised word appears in light blue font. Since the word "love" can also be used as a noun, we tested the generated results with the following phrases:

"Love is a great feeling." – "Amare è un grande sentimento." / "L'amore è un grande sentimento."

Here the first one, “amare”, was marked blue, thus recognised by the system, and in the second phrase the glossary term was not applied.

Here is another example:

“Your love is not so deep.” – “Il vostro amore non è così profondo.” / “Il tuo amore non è così profondo.”

“She is his great love.” – “È il suo grande amore.” / “Lei è il suo grande amore.”

In these two examples, the word love was not identified as a glossary word and the correct noun was used.

So, we can assume that in some cases, there are glossary systems (or tools used in addition to a glossary system) that can tackle homographs.

Although this brief experiment cannot guarantee that the results are constant and that context is taken into account in any way, it is certain that an exact-match glossary system would not be able to differentiate between the source verb “love” and its homographic noun with no additional tools or post-editing. This means that it would apply “amare” in all of the above cases, which would make the sentences ungrammatical.

From the examples above we can observe some of the ways in which glossaries used by machines work. Some of them, like the glossaries at Unbabel, work exceptionally well with exact matches and provide the equivalent of the source term without modification. However, there are others where only one form is included in the glossary and inflections are already applied in the target text, taking into account the context and distinguishing between homographs. In many cases, glossaries are created to be used for translating specialised texts. This is also the case with Unbabel. Therefore, in the next section, we will examine some topics related to specialised language and, in particular, multiword expressions, which are the type of glossary units that will be further analysed in this work.

3.4 Multiword Expressions for Special Purposes

This section focuses on Multiword Expressions (MWEs), Language for special purposes (LSP) and Multiword Expressions for Special Purposes (MWESP). Language for Special Purposes (LSP) is used in a particular field or domain and is characterised by terminology, abbreviations, and jargon that do not appear in non-specialised language, such as everyday discourse. LSP

plays a fundamental role in facilitating communication within specialised communities, enabling precise and efficient communication of complex ideas and concepts.

MWEs can be any combination of words that are frequently used together in a particular language or context. They are usually not specific to a particular field, e.g. "close a deal", "get the message" and "pay someone a visit", but they can also be used in a specialised field. In this case, we will refer to them as MWESP.

As for the definition and description of the multiword expressions, Masini (2019) claims that: "Sometimes 'MWE' is employed as an umbrella term that refers to this vast domain (as in this article); sometimes other terms are used instead, for instance, 'idiom', 'collocation', 'locution', and 'fixed expression'." The author gives a brief historical overview of the terminology used for referring to multiword expressions from the beginning of the 20th century until today. In Masini's (2019) overview, terms such as "fixed expression", "lexical phrase", "syntactic compound", "locution", "phraseme", "formulaic sequence", "idiom", "listeme", "multiword lexical units", among others, are mentioned, which again highlights the lack of consensus among scholars:

> This huge terminological variation reflects the lack of a unified tradition and of communication between different research areas, a basic uncertainty as to where in the language faculty MWEs should be found and treated, and a certain conceptual complexity regarding MWEs [...].

LSP can be challenging to understand by those unfamiliar with the terminology of a specialised field. An additional challenge can be introduced by units that are composed by two or more words, i.e. MWESP. This is because in many cases it is the combination of some common language words (rather than a particular term within the multiword unit) that assigns the whole unit to a specialised concept. Thus, a person might be able to identify the word for word meaning of the elements making up the expression but not the specialised field concept associated to the whole expression, e.g. "ring fence" and "red clause credit" from the finance domain.

Smadja (1993) explains this idea in the following way, mentioning the sailing domain:

> In addition to nontechnical collocations such as the ones presented before, domain-specific collocations are numerous. Technical jargons are often totally unintelligible for the layman. They contain a large number of technical terms. In addition, familiar words seem to be used differently. [...] Linguistically mastering a domain such as the domain of sailing thus requires more than a glossary, it requires knowledge of domain-dependent collocations.

In regard to translation of multiword units and collocations, Poddar (2013) states: "Collocations differ from language to language. So a multiword expression cannot directly be translated from one language to another conserving its inherent idiosyncrasy or metaphoric meaning".

In the context of analysing the functioning and requirements of machine and human translation, it should therefore be mentioned that there are some key differences between the translations of multiword expressions and MWESP produced by human translators and by automatic translation engines. Human translators can recognise idiomatic expressions and cultural references and use their creativity and imagination to produce translations that reflect the spirit and tone of the original text. Machine translation engines, on the other hand, may struggle to translate accurately collocations or MWESP because of the possibility to recognise only the word-for-word level of the meaning.

An additional interest presents the translation of multiword proper nouns by humans and by machines. There are several strategies to deal with this and they often have to be applied on a case-by-case basis. When it comes to multiword proper nouns used in LSP, the challenges for successful translation become even greater. These challenges were the reason for selecting the glossary units for analysis in this text. This is due to their structure and nature: multiword proper nouns used in a specialised context.

To better understand these units and the possible strategies used to translate them, a brief overview is given in the following section. It explains the characteristics of MWESP and the challenges they pose to human and machine translation. It also discusses strategies for effective translation, focusing on multiword proper nouns used in LSP.

3.4.1 Characteristics and Types of MWESP

As already mentioned above, there is no unanimous opinion among the scholars of the different areas regarding the characteristics and the existing types of MWESP. Various classifications exist and new ones continue to be presented according to the course of thought of each author or the school they belong to. Maybe the one thing that is agreed among scholars is that MWEs vary greatly: in their structures, functions and usage; they also include an enormous amount of examples throughout all languages.

For example, Poddar (2013) mentions that MWEs are very common in natural language and give fundamental knowledge about a language, however, they are also highly irregular in their nature (p.3, 2013).

A similar idea is expressed by Smadja (1993): "Collocations vary tremendously in the number of words involved, in the syntactic categories of the words, in the syntactic relations between the words, and in how rigidly the individual words are used together." According to Mendes and Antunes (2016):

> The term multiword expression (MWE) is frequently used to encompass different types of lexical sequences that present some degree of lexicalization that ranges from fully lexicalized idioms to collocations, i.e., co-occurrences between two or more words that tend to be more frequent than expected based on the frequency of each element in a corpus. This may include a diversified set of MWE, such as collocations, nominal compounds, idioms, formulae, proverbs, and light verb constructions [...].

The idea of lexicalization as a factor of classification of MWEs appears in Sag et al. (2002) as well. Sag et al. divide MWEs into two broad categories: lexicalised phrases and institutionalised phrases. The difference between the two is that the former have fully or partially idiosyncratic syntax or semantics while the latter are semantically and syntactically compositional, however, they do occur very frequently in a certain context. According to Sag et al. (2002, p. 3) lexicalised phrases can be further divided into *fixed expressions*, *semi-fixed expressions*, and *syntactically-flexible expressions*. *Fixed expressions* are "immutable expressions in English that defy conventions of grammar and compositional interpretation" and do not undergo any morphosyntactic variation or internal modification. *Semi-fixed expressions* do not change in terms of word order and composition, but present some degree of lexical variation (for example variation in inflexions, determiner and/or reflexive form). In this category these authors include non-decomposable idioms, some compound nominals and **proper nouns**, which are the focus here (see p. 65). *Syntactically-flexible expressions* present "a much wider range of syntactic variability" (Sag et al., 2002, p. 6) and can be subdivided into *verb-particle constructions* (which can be semantically idiosyncratic or compositional), *decomposable idioms* (which "tend to be syntactically flexible to some degree", *ibid.*) and *light verbs* (which are highly idiosyncratic and present total syntactic variability). *Institutionalised phrases* have compositional nature in terms of semantics and syntax, but are idiosyncratic on a statistical level, which means that the elements of the phrase tend to occur together in a certain context, such as "traffic light", as mentioned by these authors.

Masini (2019) suggests that MWEs can be generally classified into the following types: "Authors categorized them according to their (i) formal properties (degree of internal cohesion or fixity), (ii) idiomatic status [...], and (iii) function, or a combination of these." In the first option, among others, it is mentioned the above classification of Sag et al. (2002) between "institution-

alized phrases" and "lexicalized phrases". Within the second type, the classification according to the "idiomatic status", Masini (2019) mentions the "idioms of encoding" and the "idioms of decoding" of Makkai (1972) and explains that the former are "idiomatic to be encoded but less problematic to be interpreted", while the latter "must be part of the speaker's knowledge both to be encoded an to be decoded". In regard to the third type, according to the function, Masini (2019) brings again the ideas of Makkai (1972) and the difference between "lexemic idioms" and "sememic (or cultural-pragmemic) idioms": "which is basically functional in that the former identify idioms with lexical function (like phrasal verbs), whereas the latter are expressions with a pragmatic function (formulae, sayings, clichés, etc.)". Masini does not mention specifically the case of the multiword proper nouns.

As per the relationship between MWEs and the field of terminology, Baldwin and Kim (2010, p.11) comment that in terminology's vast history of research on multiword terms, focus is on the identification and classification of technical terms which are part of a specific domain and their patterns of variation. They can be MWEs and lexemes on their own. Thus, logically, the authors explain that the scope of MWEs is not the same as the scope of terminology, because in MWEs are no non-multiword units and in terminology there are no non-technical MWEs.

Laporte (2018) states that: "Forty years after the first published comprehensive classifications of sets of MWEs, the community has not reached a satisfactory consensus on large classes or on the most relevant features" and that "This uncertainty confuses computer scientists' main MWE-related activity, which is to recognise types of MWEs in texts through statistical engineering: the community does not offer a consensual definition of types of MWEs".

3.4.2 Multiword Proper Nouns

Sag et al. (2002) describe multiword proper nouns as "syntactically highly idiosyncratic". They analyse some multiword proper nouns of U.S. football teams (e.g. *the San Francisco 49ers*, *the Oakland Raiders*, Sag et al., 2002, p. 7) and present some arguments why they cannot be considered *fixed expressions* (also described as *words-with-spaces* within that source). One of them is the fact that part of the name can be omitted in an optional manner (in that case this refers to the first part of the name of the football team, which corresponds to a place or an organisation name, e.g. *the 49ers*, *ibid.*). Another argument is the topic of the determiner, which is associated to the team name always when the name occurs as a noun phrase, but not when the name occurs as a modifier in a compound name, e.g. *an/the [[(Oakland) Raiders] player]*, *ibid.* (in those cases, the determiner refers to the whole compound noun and not to the team

name).

For part of the analysis of the glossary units considered in this work will be used a method proposed by Lincoln Fernandes (2006), who analysed translations of names in children's fantasy literature. Fernandes's method is built upon the classification of strategies for translation of onomastic material by Hermans (1988).

Fernandes (2006) also follows the idea of Hermans (1988) for the division of names into two groups, regarding the necessity of translation: *conventional* and *loaded*. The translation of the first group is "unmotivated", either because of the lack of semantic load of the name, because the morphology and/or the phonology of the name does not need any adaptation, or because the name has already gained an international status. The second group, the *loaded* names, is composed of semantically loaded names, which can be separated into *expressive* and *suggestive* names or nicknames. The semantic load of expressive names is evident, and they create a link with the lexicon of the language. Such names can be, for example, *Destiny*, *Rose* or *Big Bad Wolf*. In the suggestive names, on the other hand, the semantic meaning is not so obvious, which is the case of, for example, *Voldemort*, according to Fernandes (2006).

Fernandes's (*op. cit.*) classification has ten types of *procedures* (or strategies) for the translation of names: *rendition*, *copy*, *transcription*, *substitution*, *recreation*, *deletion*, *transposition*, *phonological replacement* and *conventionality*.

The first one, can be applied in cases "when the name in a source text is enmeshed in the lexicon of that language, thus acquiring "meaning" to be rendered in the target language" (Hermans, 1988, as cited in Fernandes, 2006). This category can be represented by the following examples: the English "Wolf" translated to *Lobo* in Portuguese in *Big Bad Wolf – Lobo Mau*.

The second procedure, *copy*, is defined by transferring the name present in the source text to the target text without any changes, namely orthographic adjustments. This author explains that this practice sometimes implies a change in the pronunciation of the original name by the reader of the target text.

The procedure *transcription*, according to the classification of Fernandes (*op. cit.*), includes any kind of orthographic adaptation of the original name to the writing system of the target language. Thus, it includes transliteration and phonological transcription, for example, when the SL and the TL use different writing systems. For example, the English "Mario", the Portuguese *Mário* and the Bulgarian *Mapuo*.

The fourth procedure described by Fernandes is *substitution*. He defines this procedure as

the act of substitution of a name with another which is not related to the first one formally and/or semantically. The provided example is the substitution of the names *Harold* and *Harvey* in a sentence in the original text of one of the books of the series *Harry Potter*: "It might have been Harvey: Or Harold." with completely unrelated names *Ernesto* and *Eduardo*: "Talvez fosse Ernesto. Ou Eduardo." (Fernandes, 2002, p. 52). In this example, the translator tried to reproduce the relationship between the two names in the same sentence, i.e. the alliteration that occurs in the beginning of the two names. The substitution is motivated by the desire to improve the readability of the text by the target readers. However, when using this procedure, one should analyse the source text carefully in order not to miss any semantic dimension of the name.

The strategy named *recreation* by Fernandes (2006) is defined by using methods to create a new lexical item in a TT, similar to the ones used in a ST. The important factor in this procedure is that it aims to produce similar effects in the target text. Also, it is necessary to point out that the lexical unit does not exist in the SL or the TL. Such examples are "Heffalump" in Winnie-the-Pooh from A. Milne (1926) and *Муслон* in the Bulgarian translation of the book, made by Vera Slavova (1987).

Deletion is a strategy in which names that are apparently rather unimportant for the development of the narrative are omitted completely or partially in a target text, e.g. *Her neighbour was called John Cholmondeley – Съседът ѝ се казваше Джон.*

The procedure *addition* aims to provide additional information when it is necessary for understanding an important semantic dimension of the source name or to solve ambiguities which arise in the target language. Fernandes (2006, pp. 53-54) gives an example in which titles are added to the target proper nouns in order to assign gender to the proper noun, bearing in mind that they refer to animals and that the names of these animals have the same form for both feminine and masculine in Portuguese: *He-Beaver – Sr. Castor.*

Transposition, according to Fernandes (2006), who based this procedure on a classification by Vinay and Darbelnet (1995) in which it is defined by the keeping of the original meaning of the text but altering the word class. Fernandes (*op. cit.*) points out that this procedure often also requires structural changes in the phrase, e.g. *Animal Farm* (adjective) – *A Quinta dos Animais* (noun).

The *phonological replacement*, according to Fernandes (*op. cit.*), should not be confused with a phonological transcription in which the name in the source text is adapted to the phonology or the morphology of the target language. On the contrary, phonological replacement involves

the replacement of the name of the source text with a different name in the target text, which has phonological or morphological similarity with the source name. This strategy aims to mimic some of its phonological features of the source name, e.g. *Myrtle – Murta* (Fernandes, 2006, p. 55).

Conventionality is a procedure described by Fernandes (2006) as a procedure in which "a TL name is conventionally accepted as the translation of a particular SL name". These names are usually toponyms or names of historical figures and literary characters (e.g. *Napoli – Naples – Nápoles - Неаполь*), and they are known as exonyms.

3.5 Conclusion

An important topic when comparing human and machine translation is the translation of multiword expressions. Machine translation engines are capable of processing large amounts of text quickly and efficiently and applying sophisticated algorithms to produce text that is as close to the original as possible. However, they are currently unable to fully understand the nuances and cultural context of language that humans possess and frequently struggle with the translation of multiword expressions. Human translators can recognise idiomatic expressions and cultural references, and also use their own creativity and imagination to produce a translation that reflects the idea represented in the original text and not only the word-for-word level. MWEs are described with different names by scholars, including idioms, collocations, locutions, fixed expressions and others. MWEs also play an important role in specialised fields and facilitate precise communication.

These are the fundamentals that one needs to have in mind when creating and translating a multilingual glossary that needs to be used equally well by humans and by machine translation engines. The impossibility presented above of identifying a common typology of MWEs shows the vast number of approaches that could be taken for the analysis of any group of MWEs.

As part of the multiword expressions, multiword proper nouns pose similar challenges for analysis and translation, especially when used in a specialised field. Strategies for translation need to be applied on a case-by-case basis and their unique features and available context must be taken into account.

In order to better understand these cases, we have undertaken a thorough analysis of the units in order to identify patterns and make some recommendations on possible solutions for the translation of multiword proper nouns. The methodology of the selection and curation of the

units is explained in the next chapter.

4 Methodology

This section presents the methodology that was used to curate the glossary units. The process of curation is key to the analysis presented in Chapter 5. Chapter 4 also includes some points related to the impact of the glossaries on the translation process at Unbabel. The aim of the curation was to briefly analyse the structure of the glossary and each glossary unit, to identify translation challenges and to suggest improvements to increase the quality of the glossaries (both source and target units). This section also addresses the important aspect of data anonymisation and compliance with data protection regulations.

Below is a short summary of the steps taken in chronological order:

1. Data collection

For data collection, we were granted access to two variants of the same glossary used at Unbabel. The differences between the two variants were numerous, but one of the most important is that the first variant (G1) contained more source terms than the second variant (G2) but far fewer target terms. Here we used only the G2, as it was the final variant used by the company.

2. Selection of glossary units

To enable an in-depth analysis, a subset of the glossary units was selected. The units of general vocabulary were not considered in the selection. The selection process focused on identifying source units with slightly different structure but identical origin and usage. One of the selection criteria was to find source units that might pose a greater challenge to the translators of the glossary, especially those that might be challenging to translate for more than one target language. Taking into account the structure of the source and target languages, we chose multiword units because they are usually more difficult to translate and can have many different translation variants that change the meaning.

We looked for similar structures in the source units in order to compare them with those in the target languages and examine the translation decisions. Moreover, the creative nature of the source units, which are names of imaginary characters in video games, as well as their characteristics, promised to pose a greater challenge for translation than the other units, whose equivalents could be easily found by consulting the resources related to the domain to which

they belonged. Certainly, there are many cases where a lot of research and creativity is required to translate a particular uncommon term, but time constraints and the need for a manageable scope in terms of volume led to the final decision regarding the scope of the analysis.

Thus, only multiword source units with more than three elements were selected. This included the definite article when not in initial position.

In this way were selected 28 source English units and their corresponding 84 target units in Bulgarian, European Portuguese and Brazilian Portuguese, which made a set of 112 units.

3. Curation

The selected glossary units were subjected to an initial analysis to reveal problematic points and possible improvements. The analysis covered various aspects, including the structure of the units and orthographic considerations. One spelling error was identified and corrected in the target units. One source unit contained brackets that did not change the meaning in any way, so a variant without the brackets was considered in the further analysis.

Problematic points in the selected glossary units were identified through careful review. Since the selected units were names of video game characters, the characters mentioned in the glossary were identified in the publicly available data on the respective video games. For some glossary units, no character was identified. This was problematic in terms of the potential impact on the quality of the translation and the usability of the glossary.

4. Anonymisation of data

Given the sensitivity of the data and the requirement to comply with data protection regulations, measures were taken to anonymise the information collected. Anonymisation involved the creation and use of pseudo-examples that reflected the structure and characteristics of the original data while protecting the identity of the client. The process complied with the General Data Protection Regulation guidelines and the company's privacy policy. The handling of sensitive data was made exclusively on company-owned computers. Sharing sensitive internal or client data was strictly prohibited under the company's privacy policy.

The methodology described above provided a systematic approach to selecting and examining the glossary units, identifying problematic points, ensuring the secure handling of data and providing a stable foundation for the further analysis of the selected glossary units. By applying this methodology, we aimed to contribute to the understanding and improvement of the use of glossaries in the translation process at Unbabel.

It is important to recognise the limitations of the study. The analysis was conducted on a

limited subset of glossary units, which do not represent the full range of challenges and complexities of glossary systems. The insights and recommendations gained from this study should be interpreted in the context of the selected glossary units and may not be applicable to all glossaries used at Unbabel or in other translation environments. However, they could be useful in many cases.

In the following pages the steps mentioned above will be observed in more detail.

4.1 Selection of glossary units

This section aims at understanding better the term selection process. First, access was given to an initial version of a glossary built by Unbabel (see section 2.3.5) and to one client glossary (which was part of the language resources available to help with the translation and is described on page 76). This first glossary (G1), a working version, was used only during the internship for initial analysis, part-of-speech identification, addition of plural or singular form and for comparison with the final version (G2) which was provided later and serves as the basis for the analysis in chapter 5.

The G1 version contained 4224 English source terms. The number of target units in my working languages were 582 target units in European Portuguese, 0 target units in Bulgarian and 2481 in Brazilian Portuguese.

G2, the final version provided by Unbabel, contained 3422 source units. 1761 units were removed from the glossary immediately. These were mainly general vocabulary, most of which came from *Tickets* (1295 entries). The removal of these units was based on the fact that they were identified as general vocabulary that should not be included in glossaries, as indicated in section 2.2.2.1.

As mentioned earlier, multiword units with more than three elements were considered more challenging for translation into various languages. As a third element a definite article was accepted but not in initial position. Thus, the multiword glossary units with three or more elements (449 source units and their respective target units) were separated in another document.

From these units, a new selection was made with only client-specific units. This was done in order to identify special and challenging concepts of this client and not only multiword units belonging to this domain. In this way, 318 units were selected.

Within these multiword client-specific units, some structural patterns of the units were

identified, e.g. the names of specific chapters within a game, objects or characters. The names of characters caught our attention because of the interesting combinations of elements, the different translation possibilities, the apparent ambiguity and the fact that each one referred to a specific named entity within the game that could be important and helpful for the translation process.

From these 318 units a smaller group of 28 units was created. The structure of these 28 units is similar. They are multiword glossary units, with a proper noun in initial position, followed by a definite description. Poesio and Vieira (1997) explain that noun phrases with the definite article *the* are known as "definite descriptions". There are some differences in the structure of the definite description, so that three different substructures were identified in the source units. These structures consist of either:

1. definite article *the* and a noun
2. definite article *the* and two nouns
3. definite article *the*, an adjective and a noun.

The origin of all but one of the selected units has been identified in the column *Origin* as *Platform* and one has been identified as coming from *Client*.

Along with these 28 English source units, their translations into two different languages were analysed divided into three columns of target units: Bulgarian, European Portuguese and Brazilian Portuguese. This gives a total of 28 source units and 84 target units.

Before proceeding with the analysis of the selected glossary units, we need to take into account some additional considerations.

4.2 Problematic points in the selected glossary units

During the curation process, some problematic points were identified within the selected glossary units. These were related to the elaboration, revision and the uniformisation of the glossary, factors that can make the glossary difficult to use. It is important to reiterate that the glossaries in the company are used by both machines and a human editors. The respective limitations have already been mentioned (see sections 3.3.1 and 3.3.2), but some formal problems will be briefly addressed in this section.

A spelling error was found when analysing the Bulgarian target units. This can cause problems in the target text, as the unit with the spelling error will appear in the post-editing tool identified as coming from the glossary. Therefore, some editors may not correct the error,

either because they do not see it or because they are reluctant to correct glossary units even if they see an error. This can happen because some editors may think that the information in the glossaries is always correct. It can also be problematic when the glossary is used in reverse, i.e. when the target glossary units in Bulgarian are used as source units in a text translated from Bulgarian into English. Thus, the misspelled unit is not going to be recognised by the glossary system and the glossary will not be triggered, so the correct target unit will not appear in the target text. In the analysis from now on, we will assume a correctly spelled variant that is closest to the misspelled unit.

The same phenomenon occurs when a glossary unit contains punctuation marks, such as brackets (or a white space between two elements of a compound word). This happens in one of the English source units. In the sources, referring to the videogames in question, the same unit was found without the brackets. A possible solution is to duplicate the glossary entry, e.g. include one unit with brackets and one without them, so that the glossary system is able to identify the unit as part of a glossary in both cases. In this analysis, however, we will assume that only the variant without the brackets is available, as they do not change or alter the meaning of the ST.

In the analysis of the source units and their context in the video games, 6 of the characters have not been found. For one of them it is a problem to identify which is the character referred to in the glossary, as two characters with similar names were found and it seemed that the glossary unit is a compilation of the two. Other characters were not found at all. This may be a problem for the reasons explained in the previous sections, but also because the inclusion of these units in a glossary may in some ways seem like an unnecessary effort and waste of resources, considering that the units do not appear in any source text.

Problems with punctuation and in particular related to missing commas were also noted.

4.3 Anonymisation of Data

During the internship, after the selection of the glossary units was made and the first steps of the analysis had been taken, emerged the obligation that all data that could reveal directly or indirectly the name of the client had to be anonymised. They could not be used or otherwise passed on in any other way. Thus, all data that were being analysed had to be made anonymous using different methods but always ensuring confidentiality and protection of client information. The structure of this work and the way the analysis was presented had to be changed as well,

which made the whole process more complex and difficult to follow.

This obligation came from the mandatory implementation of the General Data Protection Regulation, which "strengthens existing rights, provides for new rights and gives citizens more control over their personal data" (European Parliament & Council of the European Union, 2016). Furthermore, the company's data policy did not allow the disclosure of sensitive internal or/and customer data.

Various methods were used to protect the data, such as: working exclusively on an Unbabel-owned computer, in and out of the office; using pseudo-examples and/or anonymised examples that replicate the characteristics of the data, such as the structure of each unit, without revealing the identity of the client.

Hence, further on in this text the elements of the glossary units analysed are represented according to their word class, and each element is marked with brackets *[ ]*, for example *[NOUN]*. The definite descriptions, which are included in the glossary units, are also marked with brackets, for example, a definite description composed by a noun and a definite article is presented as: *[[ARTICLE] + [NOUN]]*.

Another type of anonymisation is to create pseudo-examples: e.g. *Blue textile dove*. Moreover, it can additionaly be anonymised as *[ADJ-COLOUR] + [NOUN-MATERIAL] + [NOUN-SG-BIRD]*, where *ADJ* stands for *adjective*, and *NOUN-SG* for *noun-singular*. This latter type of anonymisation, which provides additional information about the class of the noun or adjective can also be used in many contexts. However, for the analysis presented here, a simpler type of anonymisation is used, where only the category is mentioned: *[ADJ] + [NOUN] + [NOUN]*. Those methods are used in the analysis of the customer data in Chapter 5.

The anonymisation process is an essential part of the work during the internship and in the preparation of this internship report. It is also important for the interpretation of the text, as the original glossary units are not revealed in this work, which may have an impact on the understanding of the concepts and the statements presented.

4.4 Resources used: in-house glossary of the client

Along with the glossary created at Unbabel, I was provided with an in-house glossary of the client to use as a language resource. The units of this internal client glossary (which referred to just one of the client's many video games) were included in both the first and second versions of the glossary analysed (referred to here as G1 and G2), and we believe that some of the

characteristics of the original in-house glossary of the client are relevant to the further conclusions of this study. Unfortunately, none of the glossary units selected for analysis in this work were found in the in-house glossary.

The original XLSX file had 766 English source units and was translated into 11 languages, including Brazilian Portuguese, but not European Portuguese or Bulgarian.

Its structure included the source units, the target units, a description, a category of the unit (an object in the game, a name of a character, etc.) and a small image of the character or object to which it referred. The latter is an important feature not only for video games localization, but also for many other fields because images can also represent concepts in dictionaries, encyclopaedias, etc., as Cabré and Sager (1999, p. 164) emphasise:

> For some types of concepts, particularly those referring to objects, images are the best mode of representation because they are easy to understand and allow users to access information starting from a general concept when other characteristics of the term in question are not known.

Given the fact that "each game creates its own unique imaginary game world" (O'Hagan & Mangiron, 2013) and that many of the units in that glossary were objects that exist exclusively in the game, it is possible to deduce that they cannot be fully understood without the image provided, an extensive description or access to the fictional world of the game.

This implies that the ideal approach is for translators to familiarise themselves enough with the game world by building up a solid knowledge of the game mechanics, storyline, characters, and so on, by applying various techniques, for example, by playing the original game. This allows them to make better decisions about word choice and tone and ensure good quality translation. As O'Hagan and Mangiron (2013:121) point out, this familiarity with the game makes the localisation process longer (and also requires in-house translators, as they have full access to information about the game before its official launch). Also, nowadays the time between the launch of the original and the localised versions is getting shorter, so this approach is not always possible. However, translators usually receive a localisation kit containing general information about the game's content, the translation project, the storyline of the game, reference materials, etc., and probably the in-house glossary sent by Unbabel's client was also part of such a localisation kit.

Stein (2018, p. 8) points out in relation to the localisation of video games that: "the localization of video games seems to be situated in between these poles [*technical and literary*

texts], as the texts are often combinations of technical and literary writing". This statement is relevant in our case, having in mind several factors, including the need to strike a balance between technical accuracy and trying to convey all the creative features of the game's elements in order to provide players with an engaging experience.

Thus, considering the similarities and the differences between technical and literary translation and localisation of video games, let us observe part of the process of creating glossaries for technical translation. As Cabré and Sager (1999, p. 105) point out, there are some principles when it comes to descriptions of terms in glossaries, namely that these descriptions "must collect all essential characteristics of each concept [...]; they must include all characteristics that are important for a complete description of the concept, even if they are not essential".

Some of these principles for creating terminological definitions are applicable to the in-house glossary of the client used for localising video games and Unbabel's glossary, but some are not. For example, glossaries at Unbabel are created per client, and often each client has various glossaries depending on their needs and area of business. This limits the use of each glossary, so including all the essential and secondary characteristics of a glossary unit is not always necessary. This is valid also for the in-house glossary of the client.

However, in some cases a clearer, more extensive and less ambiguous definition is required. The type of definitions available in the in-house glossary provided can be seen in the following Table 4.1, using pseudo-examples:

Table 4.1: Examples of definitions of the source units

en	Definition	pt	pt-br
Actually Elephant	Stylized for a real elephant	Elefante legítimo	-
Chilled Bear	Has adapted well to warmer climates	-	Urso Veranista
Crane	Decoration	-	Grou
Daisy	-	Daisy	Daisy
Kitty	Collectible	-	Kitty

Although many of the glossary entries probably do not need clearer definitions, there are some for which more context is necessary for them to be translated correctly. Some problematic points that were identified using the pseudo-examples presented above can be mentioned, such as homonymy. Bowker and Ciro (2019:46) affirm, in regard to homonymy in MT that "[...] choosing the wrong equivalent is bound to result in a nonsensical text". For example, the definition of *Actually Elephant* partly explains the concept, mainly its form, but does not refer to its purpose, function or how it differs from the real-world animal. The same can be observed in the second

example. In the third example, the word *crane*, according to the definition in the glossary of the client is an object used in the game for decoration. However, according to the definition in the *Cambridge Dictionary* this word represents various homographs, two of which are compatible with the definition *Decoration*, depending on the fictional world of the game:

> **crane**. n. (bird) a tall bird with long, thin legs and a long neck.
> **crane**. n. (machine) a tall metal structure with a long horizontal part, used for lifting and moving heavy objects.[1]

If this word was included in a glossary used in the language pair English – Portuguese, the homonymy could result in an incorrect translation, as not all Portuguese translations of *crane* are homographs. The feminine form of the bird is, in fact, a homograph of the machine, however the masculine form of the bird does not have the meaning of machine:

> **grou**. masc. n. (Ornithology) crane: any bird of the *gruidae*, large wading birds with long legs, neck and bill.[2]
> **grua**. fem. n. 1. (Ornithology) female crane. [...] 3. hoist crane, derrick.[3]

In other words, additional context is needed, for example, an image of the object, as mentioned on page 77.

In the context of providing information for translators, it is noteworthy that the majority of the terms in the in-house glossary, namely 720 out of 766 terms, are accompanied by illustrative images. In contrast, the Unbabel glossary does not contain any visual representations or hyperlinks to such representations. Consequently, it can be said that translators working on video game localisation can benefit from a richer set of resources provided, especially due to the inclusion of images. Regrettably, this is not uniformly true for translators using the Unbabel glossaries. It is important to mention that at the time of the internship the Unbabel glossary system did not include a built-in feature for incorporating images. However, it offered the possibility to include additional explanatory descriptions and hyperlinks that lead the translator to external image sources and relevant information.

Another important observation regarding the translation of proper nouns in the in-house glossary of the client is that a quick review of the proper nouns contained therein suggests that the translators who localised the game had more freedom in changing the proper nouns than the

[1] Retrieved from `https://dictionary.cambridge.org/dictionary/english/crane`
[2] Retrieved from `https://michaelis.uol.com.br/moderno-ingles/busca/portugues-ingles-moderno/grou/`
[3] Retrieved from `https://michaelis.uol.com.br/moderno-ingles/busca/portugues-ingles-moderno/grua/`

glossary translators at Unbabel. In it, there are several examples of names that vary from one language to another, such as, *Carl*, which may appear in the other languages as *Karl*, *Carlos*, *Kpuc*, *Roberto*, and so on.

This distinction is significant because it potentially affects the translation results of glossary units. On the one hand, the restriction to change proper nouns may lead to the loss of creative elements such as phonetic effects (see section 5.4), on the other hand, the decision not to change them is related to the consistency of the glossary and its usability (see section 2.3.5.4 on information on the impact of changing proper nouns in the different target glossaries).

4.5 Conclusion

In this study, we applied a systematic methodology to select and curate glossary units in order to obtain a set of units suitable for further analysis. The methodology involved several stages, including data collection, analysis of resources available (such as the customer in-house glossary), selection of glossary units, curation, and data anonymisation.

In summary, the methodology used in this study allowed for a comprehensive analysis of the glossary units selected and identified problematic points. In this way, we aim to contribute to the understanding and improvement of the use of glossaries in the translation process at Unbabel. In the following chapter, we will provide further analysis and insights into the morpho-syntactic structure of the units corresponding to multiword proper nouns, the translation challenges involved and the losses and gains in translation regarding some patterns and effects identified in the source units.

5 Data Analysis

This chapter forms the central part of this work and is dedicated to the analysis of the glossary units. It is divided in four parts. The first one consists of the morpho-syntactic analysis of the structure of the source and the target units, some strategies used for translation and the coherence in their use. In this part of the analysis we are focusing on sequences of morpho-syntactic categories in the linear order of the components. The second part consists of some considerations about orthography and punctuation of the units. On the third part there are some considerations regarding gender. The analysis then moves to another dimension, and on the fourth part the focus shifts to the creative aspects of translation, especially the translation of phonetic effects. In it we describe the phonetic effects identified in some of the source units and whether (and how) those phonetic effects were applied in the translation. Some of the effects found correspond to alliteration, onomatopoeic alliteration and rhyme. They are not exclusive and in some of the units more than one can be identified.

Due to reasons related to Unbabel's policies regarding data protection and European legislation about personal data of individuals and companies (as mentioned on page 75), the glossary units discussed in this work are made anonymous via pseudo-examples and by non-disclosure of the original data, in order to protect it. Thus, the glossary units are represented by statistical information, as well as by pseudo-examples created by keeping the relevant characteristics or structure of the original data, but without exhibiting the original glossary units.

Translation of glossary units, as any other translation, is a multi-layered linguistic process that goes beyond the transfer of literal meaning of words from one language to another. In the case of glossary units that are names of video game characters, the translation process needs to consider a diversity of facets of each unit, including the existence of phonetic effects. These effects, which include rhyme, alliteration and onomatopoeic alliteration, contribute to the unique sound properties of character names, increasing their memorability, as well as the thematic resonance within the game experience. Translating such names requires a balance between accuracy of meaning and fidelity to the creative elements and effects, challenging translators to find the proper balance between the two.

Kvillerud (1985:197 as cited in Bertills, 2003:161) points out that alliteration is commonly used as a stylistic device in children's literature and young adult fiction. In the case of the glossary units analysed they are part of video games. Although games are usually associated with children, this is not always the case of video games, which are often age restricted. In the case of the video games from which the glossary units were taken, that age varies, depending on the legislation of the country where the game is played, ranging between 13 and 16 years old. In this regard, it is important to say that the games mentioned have some similarities with children's literature and comics. This is due to the attractive, colourful graphics, which are reminiscent of illustrated children's books, as well as the character development and, in some cases, the morally instructive component. However, this does not imply that the games can or should be played by children, but it shows that its audience (which in some sources was identified as women between 30 and 50 years old) is attracted and enjoys the resemblance of the games to children's literature and comics. This resemblance can be defined as: a narrative similar to that of children's literature, including a good and a bad character; a protagonist which in some cases is an in-game character and in others is the player of the game (likewise a reader of a book); some kind of difficulty or a challenge that has to be overcome by the character (in the specific games considered in this work, this is a repetitive action, which involves the player's quick reaction or logic and gets more and more difficult as the narrative unfolds); visual effects comparable to comic books or animated movies for children; progress in the narrative, which can be positive or negative, according to the performance of the character; deliberately sought out comic effect. Logically, these similarities extend to the names of the characters (some of which are being analysed in this dissertation) and that justifies the presence of comic effects, parallelism and other phonetic and semantic effects in them.

These games are part of the so-called "casual and social game genres" (Chatfield, 2010:33) and they "target a wider audience than hardcore gamers with games playable, for example, on social networks" (O'Hagan and Mangiron, 2013:21). Or as Bernal-Merino (2014:280) defines them:

> [*Casual games are*] used by a mass non-gamer audience, typically distinguished by their simple rules. They have low production and distribution costs for developers and publishers.

These authors distinguish the above-mentioned types of games from the ones produced traditionally to be played on personal computers or game consoles in three aspects: cost, weight and time. The casual and social games are made to be played usually on a mobile phone or tablet and some have a version to be played online on a computer. Thus, they can be played in

small portions of time, for example during commute, they are lightweight in terms of electronic resources occupied by the game and usually they cost little to be produced. However, as O'Hagan and Mangiron (2013:21) assert: "Yet they can be just as engaging and popular as console games".

Regarding the characters in the games, they depend on the characteristics of each game and are inserted in the game context and narrative. They are presented using audiovisual methods. In some cases, the player participates in the narrative (as a character) and there are also in-game characters which "help" or "hinder" the task that has to be accomplished by the player. In others the player performs the action or task by interacting with the game using an in-game character, which is involved in the narrative. The games are usually divided in *episodes* and in each there may be a different background, characters and some variation of the narrative. The episodes are divided in *levels* and the progress in the game can be done after the accomplishment of a level. There are primary characters, which are present in most of the levels and secondary characters which appear sporadically in the narrative. In general, one can distinguish the "good" from the "bad" characters effortlessly, due to their facial expressions, visual attributes and/or speech characteristics, by analogy with children's literature and movies. Usually the action which the player has to perform in order to progress in the game does not vary a lot during the course of the game, however there are numerous characters, episodes and levels, which aim to make the game less repetitive and more appealing to the player. This also corresponds to one of the main goals of that type of video games, namely, to maintain a long-lasting interest of the player - idea, which is also supported by the launch of theme episodes, outfits of the characters and special levels depending on the season of the year or on specific annual celebrations.

In this sense, the goals of the translation and localization of the games should be identical to the goals of the producers of the original games. Hence, the first task of the translator of such video games should be to identify the whole spectrum of meanings incorporated, for example, in the names of the characters, in order to make an informed decision about which of these meanings must be kept in the target name or if choices need to be made.

Regarding the translation of names in fiction, Nord (2003:183) points out:

> We may safely assume, therefore, that there is no name in fiction without some kind of auctorial intention behind it, although, of course, this intention may be more obvious to the readers in one case than in another.

Therefore, in some cases, a thorough analysis of the names should be carried out in order to reveal all semantic levels and/or phonetic effects. Newmark (1988:16-17) argues that "the

greater the quantity of a language's resources (e.g. polysemy, word-play, sound-effect, metre, rhyme) expended on a text, the more difficult it is likely to be to translate, and the more worthwhile". Thus, regardless of the translation strategies used, it can be assumed, that not all of the language resources in a source text can be transferred to the target each and every time. Roderick McGillis (1996:15-16) observes that the readers can never be sure of the author's intentions because there are no absolute meanings in a text, and that it is always a question of interpretation and understanding why a specific name with its own content is given to a character. Hence, the reader (and also the translator or the player) cannot be absolutely sure of all possible meanings of a name which were thought of by the creators of the games. However, in the case of a translator, one should make an effort to discover relevant embedded meanings and effects and when not all of them can be translated, to create a hierarchy of priorities and translate the most relevant in each context, without ignoring the onomastic system of each game or group of similar games.

Considering that there are phonetic effects such as phonological parallelism (alliteration, onomatopoeic alliteration and rhyme) in 57 % (16 out of 28) of the units considered in this work, we can assume that these effects are not random, but that the authors were looking for this kind of effects. Their detailed analysis will be presented in section 5.4. Before that, however, we will take look at the surface structure of the units by considering their morpho-syntactic structure in the next section.

5.1 Morpho-syntactic structure

In this section, the morpho-syntactic structures of the English source units are examined and compared with their equivalents in the target units in Bulgarian and Portuguese. A quantitative analysis of the morpho-syntactic structures is made and the way how they are transferred to the target languages is examined. There are structures in the source language that are not expected to appear in the target languages due to contrastive characteristics of the source and target languages. In those cases different structures are expected to be found. All morpho-syntactic structures are presented in an anonymised way either by using word classes or by using pseudo-examples illustrating them.

5.1.1 Morpho-syntactic structure of the source units

This section focuses on the quantitative distribution of different morpho-syntactic structures of the selected glossary units in English and their translations into Bulgarian and Portuguese. The aim is to observe the structural variations in these languages when dealing with glossary units that follow the pattern [PROPER NOUN] + [DEFINITE DESCRIPTION]. In English, this structure is common and is used without a comma after the proper noun. In European Portuguese, however, a comma is usually placed after the proper noun to maintain grammatical correctness and natural flow, which can lead us to expect commas in Portuguese target units.

The source glossary units analysed consist of 28 entries, all of which come from the category *Platform*, with the exception of one entry which comes from the category *Client*.

As mentioned on page 74, there are three types of morpho-syntactic structures identified in the source units that can be divided into three groups.

Below is a table 5.1 which refers to the distribution of the different source structures in the selected glossary units:

Table 5.1: Morpho-Syntactic Structure of English source units

Structure of English units	Number of units	Percentage
[P.NOUN] + [[ARTICLE] + [NOUN]]	19	67.86 %
[P.NOUN] + [[ARTICLE] + [ADJ] + [NOUN]]	6	21.43 %
[P.NOUN] + [[ARTICLE] + [NOUN] + [NOUN]]	3	10.71 %

The most numerous group is the one composed by units that have a single noun in the definite description after the proper noun. This morpho-syntactic structure can be represented as: [P.NOUN] + [[ARTICLE] + [NOUN]], e.g. *Coco the Snail*.

The units of the second group have a structure that can be represented as [P.NOUN] + [[ARTICLE] + [ADJECTIVE] + [NOUN]], e.g. *Coco the Small Snail*.

In this group, we can anticipate that there will be an almost obligatory change of word order in the target units in Portuguese because of the usual word order in adjective-noun modification in Portuguese, in which the noun is followed by an adjective.

This does not apply to Bulgarian, where the usual word order is an adjective followed by a noun, as in English.

There are other source entries selected, which follow a slightly different pattern, namely, in the definite description there are two nouns. These follow the structure: [P.NOUN] + [[ARTI-

CLE] + [NOUN] + [NOUN]], e.g. *Coco the Coconut Snail*.

Considering that the units in this group contain two nouns in the definite description, one can surmise that translations into target languages that do not use this structure, such as Portuguese and Bulgarian, will contain different structures of the definite description, various translation strategies and diverse interpretations.

All these structures and pseudo-examples of the source units are included in Figure A.1 in Appendix A, together with the structures and pseudo-examples of the target units. Comparison of the structures between the source and the target units is presented in more detail in section 5.1.3 as well as some translation strategies that were used.

The **word order** of all source units is similar: a proper noun at initial position, followed by a definite description which always starts with the definite article *the* and is followed by variable element(s), a noun, an adjective and a noun or two nouns.

In regard to the **definiteness**, all source units contain a definite article in initial position of the definite description, as already shown in Table 5.1.

5.1.2 Morpho-syntactic structure of the target units

5.1.2.1 Morpho-syntactic structure of the Bulgarian target units

In this section we will analyse the morpho-syntactic structure of the Bulgarian target elements as well as some peculiarities related to word order and definiteness.

In general, when translating from English into Bulgarian, due to the structure of these two languages, the question of the gender of glossary units can be problematic in the context of MT and human post-editing of MT. Gender considerations are explored in detail in section 5.3.

The category definiteness is included because in Bulgarian, in addition to the question of gender, there is also the question of the morphological marker of definiteness (see p. 88 for a detailed description), which is agglutinated to one of the elements of the definite description (which is not always the same) and does not always have the same form.

The Bulgarian target units are organized into three morpho-syntactic structure types as shown in table 5.2.

Each target unit in the most numerous group (57.14 % of the units in this target language), consists of a proper noun and a noun with a morphological marker of definiteness. This group

Table 5.2: Structure of Bulgarian target units

Structure of Bulgarian units	Number of units	Percentage
[P.NOUN] + [NOUN + ARTICLE]	16	57.14 %
[P.NOUN] + [[ADJ. + ARTICLE] + [NOUN]]	9	32.14 %
[NOUN + ARTICLE] + [P.NOUN]	3	10.71 %

can be described as [P.NOUN] + [NOUN+ARTICLE], e.g. *Коко Охлюва*, where *Коко* is the proper noun, *Охлюв* is a noun and the final *а* is a short form of the morphological marker of definiteness in Bulgarian.

The second biggest group (32.14 % of the units) is the structure [P.NOUN] + [[ADJ. + ARTICLE] + [NOUN]], in which the proper noun is followed by a definite description containing an adjective with a morphological marker of definiteness and a noun, e.g. *Коко Кокосовия охлюв*. Here, *Коко* is a proper noun, *Кокосов* is an adjective, *ия* is the morphological marker of definiteness and *охлюв* is a noun.

The last structure present in the Bulgarian target units is [NOUN + ARTICLE] + [P.NOUN], e.g. *Охлюва Коко*, where *Охлюв* is a noun, *а* is a morphological marker of definiteness and *Коко* is a proper noun.

We can see that there are really only two different structures in the Bulgarian target units: one consisting of a proper noun and a noun, and the other consisting of a proper noun, an adjective and a noun. The first structure occurs in two variants because the word order is changed. Also, as expected, in comparison to the English source texts, the structure consisting of two nouns is not present at all.

As for the word order, we should consider how the word order of the source units is transferred to the target languages. In the Bulgarian target units, the word order of the nouns and the adjectives is not used in a consistent way, it varies.

According to Stoyanov (1983a:170):

> In modern Bulgarian, the normal word order of a word combination of a common noun and an adjective as its non-detached attribute is the one in which the adjective precedes the noun. The change of this word order is almost always perceived by Bulgarian speakers as a deviation from the word order norms of standard language or as a stylistically marked variant of the corresponding word combination [...][1]

Stoyanov (*ibid.*) gives some examples of a normal word order - "*висок връх*" [a high peak], "*велик*

[1]My translation.

поет" [a great poet], "*български език*" [Bulgarian language], and some examples of inversion - "*облак тъмен*" [cloud][dark], "*дърво високо*" [tree][high], "*вода студена*" [water][cold]. The latter examples are frequently found in Bulgarian folk songs and tales and represent a stylistically marked inversion.

Traditionally, the word order in multiword proper nouns of fictional characters in Bulgarian folklore and literature follows the mentioned "normal word order": [NOUN]+[PROPER NOUN], as illustrated by *Баба Меца* (*Grandmother Metsa*, where *Metsa* is a proper noun and also a diminutive form of *bear*), *Зайо Байо* (literally *Bunny Bayo*, the last word is also used as a diminutive of *older brother* or older man in general), *Жаба Жабурана* (*Frog Zhaburana*, the proper noun being a made-up word, derived from the noun *frog*), *Калинка Малинка* (*Ladybug Malinka*, where *Malinka* exists as a proper noun but also corresponds to a diminutive form of *raspberry*).

In the glossary units, this word order is used in only 3 out of the 28 units (11 %). The remaining 25 units have an reversed word order: [P.NOUN]+[DEFINITE DESCRIPTION], which corresponds to the word order used in the English source units and can be considered uncommon for the Bulgarian language.

Sometimes inversion also changes the focus of the semantic structure. Take, for example, a made-up name for a fictional character called *Gerard the Leopard*. If we imagine that the character is a leopard called Gerard, the translation into Bulgarian might be *Леопарда* [the leopard] *Герард* [Gerard]. If we invert the word order to *Герард Леопарда*, it can be perceived as a leopard called Gerard, but another semantic dimension also appears, namely, the character is a person called Gerard, whose nickname is Leopard, attributing to the character some typical characteristics of the animal, for example, its strength, speed or ferocious behaviour. The same is valid for English, cf. *Gerard the Leopard* and *The Leopard Gerard*, in which the second one is an animal, but the first one might be an animal or not. Hence, the source and the target units with this word order have more semantic facets and nuances if no further information about the character is provided to the players of the game. Also, we need to consider that the word order [P.NOUN] + [DEFINITE DESCRIPTION] of the Bulgarian target units conserves better the characteristics of the English source units.

As per the definite article in Bulgarian, it is considered *a formal morphological marker of definiteness* (Stoykova 2011:487 and Stoyanov 1983b:115) and not a particle, simple ending, suffix or prefix (Stoyanov 1983b:115). It comes at the end of a nominal element (noun, adjective, numeral), a complete form of a possessive pronoun or, in the case of a participle, after all

suffixes and endings. The definite article in Bulgarian can have three different functions. It can "assign an individual, or quantity definiteness, and it has a generic use as well" (Stoykova (2011):487). For each of the three grammatical genders, there are different morphological markers for expressing definiteness and one for the plural. Additionally, masculine has two different markers for definiteness for words ending in a consonant - one full and one short form. For nouns, for example, the full form is used when the noun is a subject, predicate nominative (predicative nominal), or an appositive of a singular masculine subject. In all other cases where the definite article is used with masculine nouns, it is in its short form. As for the word to which the morphological marker of definiteness is applied, in the case of a prepositive attribute, it receives the morphological marker of definiteness, and in the case of a postpositive attribute, the morphological marker of definiteness is attached to the noun (the morphological marker of definiteness is thus attached to the first element of the phrase). It is important to note that the morphological marker of definiteness is not used in proper nouns, except for some specific proper nouns and some diminutive forms. According to Stoyanov (Stoyanov 1983b:120), nicknames should always have a morphological marker of definiteness (except when used as vocatives) and masculine nicknames in the singular must be given the short form of the morphological marker of definiteness (with minor exceptions in literature). The analysed units have similar characteristics to proper nouns followed by nicknames, which means that the latter should be given the short form of the morphological marker of definiteness for the singular masculine nouns and adjectives.

It can be observed, that among the analysed glossary units there is incoherence in the use of the full and short form of the definite article in these cases.

The full form is represented by the morphological markers of definiteness *-ЪТ* and *-ЯТ* (depending on the last consonant of the word), and the short one by the markers of definiteness *-А* and *-Я*.

In 20 of the target units there are masculine nouns (one of them is an invented word, but from the suffix we can judge it is a masculine noun). In 19 of them the short form is used, but in one the full form occurs.

Another relevant factor is that the observed units contain a proper noun plus a definite description. After a brief search in the user forums linked to the games, it can be observed that these names are not always used in their complete form. Thus, in some cases, only the proper noun is used in the comments in the forums, without the definite description. Another possible suggestion would therefore be to split the units containing proper nouns plus the definite description into two units. For example, one unit containing the full form and another containing

only the proper noun. This way, the glossary system has a higher probability of recognising the glossary unit in a source text.

5.1.2.2 Morpho-syntactic structure of the European Portuguese target units

The European Portuguese target units can be divided into 6 distinct groups regarding their structure, as shown on table 5.3 below.

Table 5.3: Structure of European Portuguese target units

Structure of European Portuguese units	Number of units	Percentage
[P.NOUN] + [[ART.] + [NOUN]]	18	64.29 %
[P.NOUN] + [[ART.] + [NOUN] + [ADJ.]]	4	14.29 %
[P.NOUN] + [[ART.] + [NOUN] + [NOUN]]	3	10.71 %
[P.NOUN] + [[ART.] + [NOUN] + [PREP.+ART.] + [NOUN]]	2	7.14 %
[P.NOUN] + [NOUN]	1	3.57 %

The biggest group (64.29 % of the units) includes target units that have the structure [P.NOUN] + [[ART.] + [NOUN]], e.g. *Coco, o Caracol*. This group includes 18 units, while in all others the numbers are significantly lower.

The group [P.NOUN] + [[ART.] + [NOUN] + [ADJ.]] can be exemplified as *Coco, o Caracol Pequeno*. The next one includes the following elements: [P.NOUN] + [[ART.] + [NOUN] + [NOUN]], e.g. *Coco, o Caracol Rei*. The group [P.NOUN] + [[ART.] + [NOUN] + [PREP.+ART.] + [NOUN]] can represent by the following example: e.g. *Coco, o Caracol da Ilha*.

The next group is composed by just one multiword target unit and follows the structure [P.NOUN] + [NOUN], e.g. *Coco Caracol*. This structure in Portuguese sounds in a similar way as a proper noun and a family name of a person, which is quite different from the original English unit.

As for the word order of the nouns and adjectives in the European Portuguese target units, it differs from the word order of the source units because the normal word order of adjectives in Portuguese language is for them to occur after the noun. One of these structures mirrors that of the source and uses two nouns, as in the following pseudo-example: *Brian, o Golfinho Framboesa* - [*Brian the Raspberry Dolphin*].

There is one case where the reverse word order can be seen as problematic. This is because there are two nouns in the definite description of the English source units. In the European Portuguese target, units the Portuguese word order rule should be followed, according to which the adjective should come after the head noun, as in the example *livro novo* [book][new]. In

some cases, the inversion is possible, but it changes the meaning of the word combination, as in *homem grande*, which means "a big/fat man", and *grande homem*, which is "a great man". In one unit, therefore, this consideration is disregarded and the meaning is slightly changed (cf. for example, *Richard, the Hedgehog King*, translated as *Richard, o Ouriço Rei* and not *Richard, o Rei Ouriço*, which corresponds to the word order in the Brazilian target unit).

Regarding the use of the definite article, in 27 of the European Portuguese target units there is at least one definite article in the definite description. In 25, the definite article is present in its simple form: *o*, *a*, and in two units it is within a contraction, which also includes the preposition *de*: *do*, *dos*. In one unit, the definite article is omitted, which changes its meaning. In this case, due to the omission of the definite article, the noun becomes part of the name and can be interpreted as a surname, as in "Mario the Wolf" translated as *Mário Lobo*, for example. This shows that the inclusion of the definite article in units with the same structure and use is inconsistent. The same is valid for the Brazilian Portuguese target units, where the definite article is missing in 5 units.

5.1.2.3 Morpho-syntactic structure of the Brazilian Portuguese target units

In the Brazilian Portuguese target units, the structures vary even more compared to that of the source and the other target units. Here we have 8 different structures and we can add a separate group for the units that were not translated at all and kept in original, as you can see in the table 5.4 below:

Table 5.4: Structure of Brazilian Portuguese target units

Structure of Brazilian Portuguese units	Number of units	Percentage
[P.NOUN] + [[ART.] + [NOUN]]	12	42.86 %
[NOUN] + [P.NOUN]	3	10.71 %
[P.NOUN] + [[ART.] + [NOUN] + [PREP.] + [NOUN]]	3	10.71 %
NOT TRANSLATED	3	10.71 %
[P.NOUN] + [[ART.] + [NOUN] + [ADJ]]	2	7.14 %
[P.NOUN] + [[ART.] + [NOUN] + [NOUN]]	2	7.14 %
[[ART.] + [NOUN]] + [P.NOUN]	1	3.57 %
[NOUN] + [PREP.] + [NOUN]	1	3.57 %
[NOUN] + [ADJ]	1	3.57 %

The structures of the Brazilian Portuguese target units can be exemplified as follows:

The structure [P.NOUN] + [[ART.] + [NOUN]] was also observed in the English source units as well as in the European Portuguese target units, e.g. *Coco, o Caracol* and represents the most

numerous group with 12 target units.

The next structure [NOUN] + [P.NOUN] is a solution that cannot be found in the other two target languages and can be represented as: *Caracol Coco*. The structure of the source units is changed and the definite article is omitted, which changes the perception of the reader.

Another group with three target units is [P.NOUN] + [[ART.] + [NOUN] + [PREP.] + [NOUN]] and it can be exemplified with *Coco, o Caracol de Framboesa*. This target structure serves as a translation for the source units that include two nouns in the definite description (e.g *Coco the Raspberry Snail*).

There is also a small group of three units in which the target copies the source.

The next two groups: [P.NOUN] + [[ART.] + [NOUN] + [ADJ] and [P.NOUN] + [[ART.] + [NOUN] + [NOUN] were already exemplified above (see page 90) and have the same characteristics as mentioned.

The next three structures are represented by one target unit each. For the structure [[ART.] + [NOUN]] + [P.NOUN], the main difference from the other structures is the word order (if we compare it with the first structure described) or the inclusion of the definite article (if we compare it with the second structure described): e.g. *O Caracol Coco*.

In the units that have the structures [NOUN] + [PREP.] + [NOUN] and [NOUN] + [ADJ], there is a particular characteristic that does not occur in any other target unit, namely the omission of the proper noun: e.g. *Coco the Chocolate Snail* is translated as *Caracol de Chocolate* in the first case and as *Caracol Achocolatado* in the second. This omission can confuse the reader and does not convey the meaning of the source correctly.

All these different structures used in Portuguese illustrate the range of strategies available and the ways in which the features of the source units can be transferred into the target language. This also shows that the translator can choose between different solutions depending on the priorities he or she chooses. In some cases, however, this may also indicate a lack of coherence in the translation choices made throughout the glossary.

Regarding the word order of the Brazilian Portuguese target units, it can be said that the majority of them (19 units) follow the word order of the source units, which is [P.NOUN] + [DEFINITE DESCRIPTION]. In five of the units, the word order is inverted: [DEFINITE DESCRIPTION] + [P.NOUN].

5.1.3 Comparison of the structure of the source and target units and translation strategies

In this section we will consider the relation between the structure of the source units and the structure of target units. In table 5.5 we can observe how the structures of the English units are transferred to Bulgarian and Portuguese. For this table, the European Portuguese target terms were combined with the Brazilian Portuguese target terms. Therefore, there is only one section for all Portuguese units. However, the structures that appear only in the European Portuguese target units are coloured in yellow, the ones that appear only in the Brazilian Portuguese target units are highlighted in blue and the structures that appear in both variants are marked with green.

Table 5.5: Structure Comparison - Source and Target Units

English	Bulgarian	Portuguese
P.NOUN + [ART + NOUN]	P.NOUN + [NOUN + ART]	P.NOUN + [ART + NOUN]
		P.NOUN + NOUN
		NOUN + P.NOUN
	[NOUN + ART] + P.NOUN	[ART + NOUN] + P.NOUN
		NOUN + ADJ
		NOT TRANSLATED
P.NOUN + [ART + ADJ + NOUN]	P.NOUN + [[ADJ + ART] + NOUN]	P.NOUN + [ART + NOUN + ADJ]
		NOT TRANSLATED
P.NOUN + [ART + NOUN + NOUN]	[NOUN + ART] + P.NOUN	P.NOUN + [ART + NOUN + NOUN]
		P.NOUN + [ART + NOUN + PREP + ART + NOUN]
		P.NOUN + [ART + NOUN + ADJ]
		P.NOUN + [ART + NOUN + PREP + NOUN]
		NOUN + [PREP + NOUN]

It should be noted that the first group, which contains a single noun in the definite description, is translated with a very similar structure in Bulgarian and Portuguese. However, there are differences in the word order of the target units, which are explained in more detail in the following sections. In the Portuguese units there is a high variability of the structures which comes mainly from the Brazilian Portuguese target units (see also tables 5.3 and 5.4).

The second and third source groups, i.e. those containing a combination of an adjective and a noun or of two nouns in the definite description, are translated into Bulgarian with a single structure: a definite description consisting of an adjective with a morphological marker of definiteness and a noun.

In the Bulgarian target units, the source structure that was translated using two different structures is the: [P.NOUN] + [ART + NOUN]. However, the difference between those two is minimal because it has to do only with the word order. In Portuguese this same source structure, along with the structure [P.NOUN] + [ART + NOUN + NOUN] has the highest variation - 5

structures used for each of those source groups.

Regarding the latter, namely the structure [P.NOUN] + [ART + NOUN + NOUN], it should be noted that it is unusual in Portuguese and could sound unnatural if used. The wide variation of structures provided by the translators of the target units shows their willingness to find a suitable alternative for this type of structure, keeping the characteristics and meaning of the source while avoiding structures that do not exist or are not natural to native speakers.

As mentioned earlier, in Figure A.1 on page 119 there is a map of the structures found in our data with pseudo-examples where we can observe the cases where the same structure was used in the target units to translate two different structures of the source units.

Regarding the different strategies used by the translators, when dealing with proper nouns, Nord (2003:182-183) argues that:

> [...] looking at translated texts we find that translators do all sorts of things with proper names: non-translation [...], non-translation that leads to a different pronunciation in the target language [...], transcription or transliteration from non-Latin alphabets [...], morphological adaptation to the target language [...], cultural adaptation [...], substitution [...] and so on. It is interesting to note, moreover, that translators do not always use the same techniques with all the proper names of a particular text they are translating.

The translation strategies identified are based on the classification of Fernandes (2006), which is explained in detail on page 66. This classification was chosen because it was designed to be applied to the translation of proper nouns, which is the main feature of the glossary units considered in our work.

Regarding the strategies used in translating to Bulgarian, in most of the units there is a combination of two strategies. The combination varies, but the most common combination is *transcription* (of the proper noun in the unit) plus *rendition* (of the head noun in the definite description). This combination is observed in 27 of 28 units.

Rendition is used to translate the nouns in the definite descriptions into Bulgarian, since these have lexical meaning in the target language, e.g. *Coco the Snail – Коко Охлюва*. With regard to the proper nouns, *transcription* is used, *Coco – Коко*. Here it is necessary to point out that Fernandes (2006) includes phonological transcription and also transliteration in the definition of the *transcription* procedure. Transliteration, in the case of different scripts of the SL and TL (e.g. Latin and Cyrillic), refers to the process of transferring the proper noun from one script to another, letter by letter. In phonological transcription, the proper noun is transferred according

to the sounds it contains, which may differ greatly from the spelling - *Esperança – Есперанса* (transliteration) – *Шпрансъ* (phonetic transcription). That is why we can additionally specify that in all units except one, a phonological transcription was used, whereas transliteration was used in only one. This is important because among the source units the exact same proper noun is used twice, but in one of the cases it is phonologically transcribed in the target unit and in the other it is transliterated. Thus, the two target proper nouns differ from each other, but the source proper noun is the same, which results in a lack of consistency in the translation of the glossary.

In one unit *transcription* and *conventionality* is used. As in the cases above, *transcription* is used for the proper noun. For the definite description *conventionality* is used. This case is observed in a unit that contains a reference to a character of North American folklore which has an equivalent in many languages, including a Bulgarian one, e.g. *Tony the Big Foot – Тони Голямата стъпка.*

And finally, in one unit *transcription* is used not only for the proper noun but also for the definite description, probably because it refers to an animal from North American folklore that does not exist in Bulgarian and therefore the translator decided simply to transcribe the noun phonologically instead of trying to create a target noun based on the meaning of the source: e.g. *Uli the Urayuli – Ули Уралолито.*

With regard to the strategies used in the Portuguese translation, four strategies can be identified: *copy*, *rendition*, *transcription* and *recreation*, according to the classification of Fernandes (2006).

With regard to the proper nouns, most of them are directly transferred from the source to the target language without any adaptation or change, which corresponds to the strategy *copy*. This is acceptable for proper nouns that are of foreign origin and do not have an equivalent in Portuguese (e.g. *Ethan the Elephant* and *Ethan, o Elefante*).

For two proper nouns, a *transcription* is made by adding diacritics to one of the proper nouns and reducing a double consonant to a single one in another, thus adapting them to their orthographic form in Portuguese (e.g. *Gabriella* becomes *Gabriela*). In the case of double consonants, the adaptation of proper nouns is not motivated because the source proper nouns are easily understood by the target reader. Moreover, among the glossary units selected, there are other proper nouns that are more difficult for a Portuguese speaker to read or pronounce and have not been transcribed into Portuguese but copied directly into the target unit.

In the case of the diacritics, however, it is important to add them according to the Portuguese

norm. This is because otherwise they cause strangeness for the reader (cf. *Mario the Rabbit – Mario o Coelho* and *Mário o Coelho*).

In summary, in the case of proper nouns that are homographic in English and Portuguese, and in those cases where the difference is a double consonant, the adaptation is not necessary, as the proper noun could be understood as something foreign or exotic, but does not cause problems in readability. However, in the case of proper nouns where the difference is in the diacritics, these must be inserted according to the Portuguese norm. This is necessary because the absence of the diacritics not only makes it difficult to read, but could also be perceived as a typing error.

It is important to mention that the translator should follow his or her decisions in a consistent way.

For one of the proper nouns the strategy *rendition* is used because the proper noun of the character is also an adjective with its own meaning enmeshed in the vocabulary of the language. Thus, the translator chose to use the corresponding adjective in Portuguese, e.g. *Biggie the Cockroach*, *Barata Grandinha*. Also, in this case the definite article was omitted completely and the structure was changed, as shown in the example.

For two other proper nouns, *recreation* is used. In these cases the target units were recreated in Portuguese using similar strategies to the ones used for the creation of the source units. For example, the first one is composed by a noun, referring to a type of sweet and a suffix used usually with female proper nouns. It is translated using the corresponding noun in Portuguese with the addition of another suffix for female proper nouns, which is more common in Brazilian Portuguese, e.g. *Biscuitine – Biscoitina*. In this, the translations to European and Brazilian Portuguese differ from each other, because they use different suffixes, e.g. *Biscoitina – Biscotinela*. The second case is an adjective in English which is related to a type of mollusc and functions also a female proper noun. The Portuguese translation refers to the same kind of mollusc. It can also be understood as a proper noun in Portuguese, but it is not commonly used.

For one unit, *recreation* is used for the definite description because the noun, which stands for a mythical creature in North American folklore, is relatively unknown internationally and there is no equivalent in European Portuguese. In this specific case, according to *Merriam-Webster Dictionary Online*, the source noun is a complex word composed of two other nouns (denoting animals) via agglutination of the first half of one noun with the second half of the other and omission of the other halves. The translator decided to create a new noun in Portuguese by using the same method, but with the corresponding nouns in Portuguese. .

Three of the definite descriptions are not translated at all, so the strategy *copy* is used for

both the proper noun and the definite description.

Thus, except for the mentioned cases, *rendition* is used for all other definite descriptions of the source units, e.g. *Coco the Snail – Coco, o Caracol*.

Overall, we can conclude that the translators used different morpho-syntactic structures and translation strategies for the Bulgarian and Portuguese target units and applied them on a case-by-case basis in order to preserve some of the main features of the video game characters' names. In a few cases, however, their translation decisions were not coherent throughout the glossary.

5.2 Orthography and punctuation

5.2.1 Orthography and punctuation of the source units

Punctuation is almost identical in all source units, with the exception of the punctuation used in two units that contain brackets or quotation marks, e.g. *Coco (the Snail)* and *Coco "The Snail King"*; the remaining units have no punctuation.

As far as capitalisation is concerned, the following capitalisation applies to most units: first capital letter of the noun and first capital letter of the elements of the definite description, except for the definite article: e.g. *Coco the Snail*. However, there are five units in which only the proper noun begins with a capital letter, e.g. *Coco the snail*.

5.2.2 Orthography and punctuation of the target units

5.2.2.1 Bulgarian target units

First of all, having in mind that Bulgarian uses the Cyrillic script, it is important to mention that in the target units all proper nouns are phonetically transcribed to Bulgarian. All definite descriptions were translated.

The Bulgarian target units adopt the punctuation used in all English source units. That is why we find one unit with brackets and another with quotation marks, e.g. *Коко (охлюва)* and *Коко "Кокосовия крал"* just like in the source units.

Although the capital letters used in the target units are not a fundamental part of the translation, it is relevant to point out that they have not been used consistently. According to the Bulgarian rule for capitalising names, the proper noun should be capitalised and in the case

of a nickname consisting of more than two elements, only the first letter of the first word is capitalised. Thus, in the case of a combination of a proper noun and a nickname, both should be capitalised. In 24 units, capitalisation is used correctly, according to these rules, but in 4, the proper noun is capitalised but not the following definite description.

5.2.2.2 Portuguese target units

The analysis of the structure of the glossary units in section 5.1 revealed a common pattern: all units contain a definite description. In this context, the definite description provides additional information and functions as an epithet. To better understand this, it is worth noting that as stated in *Dicionário Etimológico da Língua Portuguesa* (Machado, 2003), the Greek word "epitheyon" means an adjective that adds a particular designation to a name. Thus, it is a word or phrase that indicates a characteristic or quality of a person or a thing for reasons of clarity or simply to enrich the noun, e.g. *Margaret Thatcher, a Dama de Ferro.* The expression that starts with a definite article, from a syntactic point of view is equivalent to an appositive. That is why those should be written without a hyphen, but with a comma (*Dicionário Prático Ilustrado*, 1959), as shown in the example above.

In addition, Cunha and Cintra (2017:664) point out that:

> To conclude our remarks, we must emphasise the following:
>
> a) any clause or part of a clause of purely explanatory value is pronounced between pauses; therefore they are isolated by commas in writing; [...][2]

Thus, in the case of the glossary units considered in this work it is necessary to have a comma between the proper noun and the definite description in Portuguese. This is so because the definite description consists of additional information that highlights the prominent feature of the entities. Hence, the target units which have the necessary comma, correspond to the above-mentioned rule, while the rest of the units do not. However, in the target units analysed in European Portuguese, only one unit has a comma, one has brackets isolating the definite description, and another uses quotation marks at the same places. Here are the different graphic representations used in the target units: *Coco o Caracol Rei*; *Coco, o Caracol Rei*; *Coco (o Caracol Rei)*; *Coco "o Caracol Rei"*.

In the Brazilian Portuguese target units, 12 units contain a comma before the definite description and the remaining 16 units do not, in one unit there are brackets isolating the definite

[2]My translation

description from the proper noun.

In regard to capitalisation, since we are dealing with names, we can assume that all main elements in the target units (with the exception of articles and prepositions) begin with a capital letter. The proper nouns in all units start with a capital letter. However, the capitalisation of the definite description is not consistently applied in the Portuguese target units. In European Portuguese, the definite descriptions of four units contain no capital letters at all, and in Brazilian Portuguese units this number amounts to 14 units.

5.3 Gender considerations

5.3.1 Gender of the source units

Another aspect of the source units that must be considered in this analysis is the gender of the names. Bertills (2003:99) points out that the consideration of gender, when translating proper nouns, is pertinent because it might be fundamental for the interpretation of the story and the character, and also that mistranslations related to gender may have serious implications on the target text. So in our case, where proper nouns are followed by a definite description, gender can have an impact on various options and aspects of agreement.

Due to the nature of the English language, the question of gender marking does not occur in the source units. However, when translating into Bulgarian or Portuguese, this question arises. Therefore, the translator must focus on this topic and should have access to the necessary context to identify the gender of the character reffered by each unit. This is important because this information affects the head noun in the definite description, and may even influence the decision on the selection of a proper noun. (Substitution of the proper nouns is not a permitted strategy in the glossaries in Unbabel, but it is allowed in other contexts, such as in the localisation of games, as mentioned in section 4.4.)

Considering the inherent gender of nouns and adjectives in some languages, such as Bulgarian and Portuguese, and the gender that can be assigned to conventional proper nouns, we have identified some problematic points regarding gender, which are mentioned below. One of them is the fact that gender-neutral proper nouns, when used in conjunction with a definite description, can be perceived as the opposite gender of that of the character of the video game, due to the inherent gender of the head noun in the definite description for instance. It is therefore important to address these difficulties in translation and to understand whether the translators have made

coherent decisions and, if not, what additional context or guidelines they need to execute their task as accurately as possible.

Thus, an attempt to identify the gender of all selected source units was made. This was performed in two steps.

The first step consisted in identifying the gender of the names using the information available in the glossary and general knowledge about the gender of some conventional English proper nouns. However, the existence of uncommon proper nouns, where gender assignment is difficult posed a problem. According to Bertills (2003:98): "As conventional first names in general are highly gender specific, the more a name is involved in language and the lexicon, the more gender-neutral it will be".

Therefore, in this step, 15 units were classified as "masculine" based on the conventional gender of the proper noun, which was identified as typically masculine. The remaining names were classified as follows: 6 feminine and 7 gender-neutral. The latter were common nouns that were used as proper nouns that could be both feminine and masculine (see Table 5.6).

The second step consisted of a brief research in the publicly available information on the Internet about the characters of the games and their gender as shown in the table below.

Table 5.6: Identification of the gender of the source units - Results

Step	Feminine	Masculine	Gender-neutral	Not found
1. Assignment based on proper noun	6	15	7	N/A
2. Research in public sources	6	16	0	6

This process aimed to confirm the gender of the source units proposed in the first step, namely the assignment according to the gender of the proper noun.

The numbers in the table above might look similar to each other, however, there was a match between the assigned genders and the confirmed ones in only 18 cases as shown in Table 5.7.

Table 5.7: Results regarding the gender of the source units

Action	Feminine	Masculine	Not identified
Match of the gender between Step 1 and 2	4	14	0
Gender-neutral units revealed as feminine or masculine	2	2	0
Not confirmed	2	1	3
Final decision	**8**	**17**	**3**

Four units that were previously marked as gender-neutral could be identified as feminine or masculine upon the research, based on the information about the character.

Three units that were identified as masculine or feminine before the research, were not found at all. In those cases, we preferred to use the previously assumed gender.

Three other units that contained gender neutral proper nouns were not found in the search. This was problematic because the gender assigned with the translation might not match that of the characters in the game.

However, apart from these 3 units for which we had no conclusive proof, we can mention that 17 units can be considered masculine and 8 feminine.

At this point it is important to mention that, given the fact that the glossary provided did not contain detailed descriptions and images of the source units, the research described was necessary. This was a considerably time-consuming task, which can be eliminated by providing more context to the glossary translators. The lack of context can also lead to incorrect translations and the assignment of an incorrect gender.

5.3.2 Gender of the target units

5.3.2.1 Gender of the Bulgarian target units

The lack of a morphologically marked gender in English nouns and adjectives makes the assignment of a gender more complex when translating into a language where there is morphological gender, such as Portuguese and Bulgarian.

Before looking at how the gender of each character was transferred to the target language, it is relevant to mention that in Bulgarian there are three grammatical genders: feminine, masculine and neuter. Therefore, some of the units identified as masculine or feminine may be translated as neuter.

An important factor in assigning a gender to the names of the characters is the grammatical gender of the head noun in the definite description of each target unit. Another factor is that some of the English proper nouns included in the source units are not associated with a gender in the target language as they are not used in the target culture, and are either perceived as unisex proper nouns or associated with a gender other than the one in the game. As an example, consider the proper noun Denize, which is feminine in English, but in Bulgarian exists only as the masculine Denis. Bertills (2003:98) points out that "[g]iven the fact of general linguistic rules, the gender of an anthropomorphic character could be considered by the linguistic gender marks in the name forms". Therefore, a combination of this proper noun with a masculine noun

is perceived as a masculine name, which does not correspond to the feminine gender of the video game character. And as Bertills (2003:225) elaborates "[...] ignoring the gender suggestive characteristics of the name may easily result in rather misleading equivalents which in turn will suggest completely new characteristics to the name-bearer".

It can be noted that in 16 of the units there is a match between the gender of the head noun in the definite description and the gender of the proper noun in the Bulgarian target unit. Thus, the proper noun and the noun are both either feminine or masculine. In 12 units there is a mix, for example a feminine proper noun with a masculine noun or vice versa and in two units, masculine proper nouns are combined with neuter nouns.

In these mixed cases, the gender of the **proper noun** and not the grammatical gender of the head noun in the definite description is defining for the perception of the gender of the character.

Therefore, it can be said that the 17 masculine source units (as previously identified in Table 5.7) are all translated as masculine in Bulgarian (and in 6 of those cases there are feminine or neuter nouns and adjectives combined with the masculine proper nouns).

For the 8 feminine source units, there are 5 feminine Bulgarian target units and 3 masculine (see Table 5.8 below). For the masculine units, the issue comes both from the nouns and the proper nouns: the proper nouns are either perceived as masculine in Bulgarian or derive from common nouns and are too unusual, which makes it difficult to assign them a gender; on the other hand, the grammatical gender of the head noun in the definite description is masculine. Thus, these three units are perceived as masculine.

The three source units, for which no context was found and were not identified with any gender were all translated as masculine and for them, the same observations apply as above: gender-neutral proper noun and a masculine noun perceived as masculine.

Table 5.8: Comparison of the gender of the English and Bulgarian units

English		Bulgarian	
Masculine	17	Masculine	17
Feminine	8	Feminine	5
		Masculine	3
Not identified	3	Masculine	3

5.3.2.2 Gender of the Portuguese target units

In Portuguese there are two grammatical genders, feminine and masculine. Looking at the selected units we can say that, as we have seen with the Bulgarian target units, when translating

nouns or adjectives from English into Portuguese, these will have a morphologically marked grammatical gender. With conventional proper nouns, on the other hand, the reader will make gender associations depending on the gender traditionally attributed to a specific name. In the case of non-conventional proper nouns incorporated in the lexicon of the target language, gender associations can be more difficult to be made, as Bertills (2003:225) mentions:

> As I have already pointed out, for semantically loaded names referring to fantasy characters, it is more difficult to determine the gender of the referent on the basis of the name - and it is not always relevant. Weighing the source name against their translation, the connotations to a specific gender are clearly relevant in a couple of illuminating examples. Although the gender of the name-bearer is usually determined in the context, regarding fantasy characters, it is still not always completely clear.

As Bertilis (*ibid.*) notes, the gender of fictional characters may not always be relevant. In our case, however, it is important because the characters will appear on an image next to the proper noun and mistranslation will confuse the player about the character and the narrative. As for the lack of context, in some cases none could indeed be found and no access to the games was granted.

According to Cunha and Cintra (2017:202), the gender of a noun cannot generally be deduced, from either its meaning or its ending. These authors have established some rules based on the usual assignment of a particular gender associated with a certain meaning or ending, which they describe in detail. They also state that in Portuguese the masculine form of nouns is the unmarked one. It can therefore be assumed that in the glossary, the masculine form of the nouns was used in some of the target units to identify only the species of animal to which the character belongs and not to indicate its gender.

When referring to people and animals, the feminine form of a noun is produced in two ways: using the same root as the masculine noun and adding inflexions (e.g. *gat<u>o</u>* - a *male cat* and *gat<u>a</u>* a *female cat*); using a different root (e.g. *marido* - *husband* and *mulher* - *wife*).

Also, there are invariable nouns, divided into three groups:

1. *Substantivos sobrecomuns*, which have only one gender in order to designate persons from the two genders, e.g. *a criança* - *the child* (either female or male).

2. *Substantivos comuns de dois géneros*, which have only one form of the noun, but distinguish between feminine and masculine by the form of the definite article (e.g. *o artista* - the male artist and *a artista* - the female artist).

3. *Substantivos epicenos* - these are nouns signifying animals, which have only one grammatical gender to designate both sexes, for example, *a cobra* means both the female snake and the male snake. However, in the latter case, there is also another way to specify the sex of the animal as Lindley Cintra and Celso Cunha (2017:209) point out, by adding the words *macho* [male] or *fêmea* [female] to the noun as in *crocodilo macho* [male crocodile], *crocodilo fêmea* [female crocodile], *o macho ou a fêmea do jacaré* [the male or the female alligator].

Adjectives in Portuguese usually have two forms - a feminine and a masculine, and they assume the gender of the noun they are syntactically linked to, according to Cunha and Cintra (2017:265). However, in some cases there is a single form, used both with feminine and masculine nouns by changing only the form of the definite article of the noun. In the examples present among the glossary units, the three units in which adjectives occur, they are of that type and they have a single form for feminine and masculine. In view of this, in the specific case of the glossary units analysed, the translation of the adjectives did not raise problems with regard to the gender marking.

17 units combine masculine proper nouns with feminine nouns or feminine proper nouns with masculine nouns in the definite description and 11 units have a proper noun and a noun with the same gender. The gender to which the proper noun is associated in the target unit leads to the attribution of a gender to the character. Thus, the 17 masculine source units defined as such earlier (see page 101) are all translated as masculine and the 9 feminine are translated as 6 feminine and 3 containing a masculine noun and a proper noun which can be both feminine or masculine.

As explained above, in order to assign a feminine or masculine gender to nouns denoting animals that do not have a feminine or masculine form of their own, one must add another noun to disambiguate the gender, i.e. [ANIMAL] *fêmea* (female), [ANIMAL] *macho* (male), but in the observed units no such disambiguation was made. The reader's attention is thus focused on the type of animal designated and not on its gender. On the one hand, this approach can be effective in making the proper nouns more concise and clear, considering that they are used in a video game where speed of action and reading is important. Also, short proper nouns are easier to remember. In the original English proper nouns, there is no way to determine the gender of the character other than the picture of the character. On the other hand, it can be relevant for identifying characters in the game because in most cases they have an appearance that indicates its gender. Gender can be relevant in cases where the same species of animal is represented by two different characters with different genders. So if the gender is not marked in some way, as it is the case in English, apart from the assumed gender of the proper noun, it can be easier to

think about the character and assign a gender to it after looking at its image within the game.

Table 5.9: Comparison of the gender of the English and Portuguese units

English		European Portuguese		Brazilian Portuguese	
Masculine	17	Masculine	17	Masculine	15
				Not translated or translated partially	2
Feminine	8	Feminine	6	Feminine	6
		Masculine	2	Not translated or translated partially	2
Not identified	3	Masculine	3	Masculine	3

A pattern can be observed: the proper nouns that are ambiguous regarding the gender, are left untranslated in Brazilian Portuguese, as two of the three untranslated units contain unisex proper nouns. We can speculate that those units were particularly difficult for the translator, who took the decision to leave them in their original form in English.

5.3.2.3 Final considerations on gender

Gender plays a crucial role in ensuring linguistic accuracy and coherence, especially in mixed media such as video games. The inherent differences between languages like Bulgarian, Portuguese and English require careful attention during the translation process.

The absence of grammatical gender in English contrasts with the obligatory gender marking in Bulgarian and Portuguese. This contrast requires careful assessment of the context to determine the appropriate gender assignment.

Proper nouns with definite descriptions are particularly complex. Here, the challenge becomes even greater when attempting to translate proper nouns of characters which include also nouns and adjectives that have inherently different genders in the target language. Harmonising these elements requires linguistic skills, cultural understanding, along with taking context into account. This context can be provided to the translator before starting the translation process: the source glossary should thus include some or all of the following elements: a detailed description of the character and an image or a link for additional information. Many of the gender issues are related to the fact that the game characters do not exist in the real world (see section 4.4), that is why access to more context is fundamental. Since these characters belong to the universe of a specific game, the translator lacks knowledge about them from real life and has to acquire this knowledge from the game context and the related information. Without the aforementioned

information regarding the context, it might be impossible to make some decisions regarding the translation. In our case, some of the characters are animals about which there is some real-life knowledge, but they may also be entirely fictional characters. Take, for example, the name of a made-up character called *Ally the Rose Bug*. This is an imaginary light green male bug with a moustache and an elegant coat who disperses multicolored roses through his umbrella shaped like a rose, which the player must catch in a certain order to win the game. In other cases, access to the context is easy or not even necessary, because the word is part of the imagination and knowledge about real life. Here, however, without a description or a picture, it would be very difficult to guess these specific features of the character in order to translate them in an appropriate way that does not cause strangeness to the player. Thus, it becomes apparent that the choices for translating these characters cannot be made without context as they can be ambiguous.

Given these observations, it is clear that a thorough understanding of gender characteristics and context is essential for the accurate translation of video game character names from English into Bulgarian or Portuguese. Therefore, the translation process requires a comprehensive approach that includes a thorough analysis of the context and the source units. That is why, we highlight the need to apply a multi-dimensional analysis of glossary units in order to achieve optimal translation results. One of those dimensions is the phonetic effects that are present in the source units.

5.4 Phonetic Effects

There is a need for a full understanding of the phonetic effects present within the source units and their role in maintaining the intended impact of each name within the narrative. Thus, in order to explore further on the characteristics of the glossary units, it should be mentioned the existence of *parallelism* in the observed glossary units. According to *The Routledge Dictionary of English Language Studies* (Pearce, 2007:136) *parallelism* is "[a] stylistic device involving prominent patterns of repetition at the level of sound, grammatical structure or meaning"

According to the same source (Pearce, 2007:136) phonological parallelism is a repetition of identical or similar sounds and can be divided in four main types: "*alliteration* (the repetition of word initial consonants), *assonance* (the repetition of similar vowel sounds), *rhyme* (the repetition of similar syllables) and *metre* (the repetition of rhythmic patterns)". Phonological parallelism was found in 16 of the 28 glossary units considered in this work.

Rhyme, according to Pearce (2007:158) is "[a] matching of sounds at the end of words (e.g.

*b**an**, pl**an**, saucep**an***)" and there is a distinction between "full rhymes" and "half-rhymes". The former are composed by a sound correspondence between final vowels or final vowel and final consonant sequences and the latter concern a phonetic closeness between the final consonants or final vowels in words and not an exact correspondence. Pearce (2007:158) points out that rhyme occurs not only in literature but rather "[i]t is also a common feature of song lyrics, slogans of various sorts and language play (particularly by children)".

In three units another phonetic effect was identified, namely onomatopoeic alliteration. While the *Routledge Dictionary of Language and Linguistics* (Bussmann, 1996:836) defines onomatopoeia as "[t]he formation of words through the imitation of sounds from nature [...]", Pearce (2007:134) considers an onomatopoeia "[a] type of sound symbolism in which the phonetic form of a word resembles the sound made by the word's referent [...]". From these definitions it becomes clear that the creation of onomatopoeic words is not arbitrary, but aims at the resemblance of the sound made by the designated thing or event. Bussmann (1996:836) also mentions that "[t]he same sound may be represented differently in other languages, e.g. *cock-a-doodle-doo* is *kikeriki* in German and *cocorico* in French". As a consequence, when a translator deals with an onomatopoeia or onomatopoeic alliteration, one possible approach is to identify the existence or non-existence of an equivalent onomatopoeia in the target language.

It is important to mention, that some sources, such as the *Routledge Dictionary of Language and Linguistics* (Bussmann, 1996:42) define the term alliteration as a "[r]epetition of homophonous accented, syllable-initial phonemes, as in *house and home*, *cash and carry*, *tea for two*, usually for stylistic or poetic effect". In this definition there is no division of consonants and vowels, regarding syllable-initial phonemes, unlike the definition of Pearce (2007:136), who considers the repetition of vowels an assonance, and the repetition of consonants – alliteration. According to Pearce (2007:10) alliteration is a "repetition of the same consonants in close proximity: *C*arrie *c*aught a *c*rab in a *c*opper *k*ettle". Also, that alliteration "can be used for emphasis, and to forge links between elements" in popular idioms, advertising and political slogans, newspaper headlines, tongue-twisters and also in verse for children, where it is used usually with a comic effect. In this work we will follow Bussman's (1996) definition regarding alliteration.

Regarding phonological parallelism, we found rhyme in 2 of the 16 units in which phonological parallelism was identified, alliteration in 15 of the 16 units and onomatopoeic alliteration in 3 of the 16 units.

Bertills (2003:161) mentions that the utilization of alliteration and rhyme in name formation can be traced back to the nursery rhymes, the phonetic aspects and playfulness in them. As

Maclean et al. (1987:255) point out, there is a strong relationship between the knowledge of nursery rhymes and the development of phonological skills in children. Thus, alliteration and rhyme are common and important stylistic devices in children's literature. As already mentioned (see p. 82), the video games containing the analysed glossary units analysed have some similarities with children's literature.

5.4.1 Phonetic effects in the source units

In the following section some observations regarding the phonetic effects in the selected source units of the glossary will be presented, as well as some pseudo-examples.

In 11 of the units in which alliteration occurs the definite description is composed of a single noun. In them, the alliteration effect appears between the first sounds in the proper noun and in the head noun in the definite description. In five units the first two sounds are repeated (e.g. **Ty**ler the **Ti**gress) and in the rest, just the first sound (e.g. **M**ary the **M**ouse).

In one unit the definite description has the following structure [ADJ. + ART. + NOUN] and the alliteration appears between the first two sounds in the proper noun and in the adjective, for example **Po**rter the **Po**lynesian Snake.

In three units the structure [NOUN + ART. + NOUN] occurs. In one of them the alliteration appears in the first two sounds of the proper noun and in the second noun of the definite description, as in **Ro**bert the Chocolate **Ro**bin. In the other unit alliteration appears in the initial sound of all elements, e.g. **D**arcy **th**e **Do**nkey **Do**ll. In one unit the alliteration only involves the first sound of the proper noun and the first sound of the second noun.

In this selection of 16 units, three cases of onomatopoeic alliteration can be found. In two of them the proper noun of a character, which is a bird, may be considered similar to a sound produced by that type of bird. In this unit, the effect is achieved by the alliteration of the phonemes [p], [r]. The other animal is a mammal, and the combination of the elements in its name can also be considered onomatopoeic to the sound produced by that animal. This effect is achieved by an alliteration of phonemes [d] and [ð] in an initial position combined with clustering of sounds like [əʊ], [ɒ], [ks], [t] and [g] in the middle or in the end of the elements.

Thus, in these three units the onomatopoeic alliteration characteristics are found when looking at the sounds of the unit as a whole. Separately, the elements of the units do not present onomatopoeic characteristics.

In two units rhyme can be found, and in one of them the number of syllables in the words

that rhyme is equal (two syllables in each element). The rhyme is obtained by the combination of a proper noun and a noun, which have similar sounding endings, for example, Be**tty** the Ki**tty**.

5.4.2 Phonetic effects in the target units

In the following paragraphs the phonetic effects observed by the source units will be considered along with their translations to the selected target languages – Bulgarian and Portuguese (European and Brazilian).

Table 5.10 represents the distribution of alliteration, onomatopoeic alliteration and rhyme in the source units analysed and in their corresponding target units in Bulgarian, European Portuguese and Brazilian Portuguese.

Table 5.10: Existence of phonetic effects in source and target units

	Alliteration				Onomatopoeic alliteration				Rhyme			
Index	EN	BG	EU-PT	BR-PT	EN	BG	EU-PT	BR-PT	EN	BG	EU-PT	BR-PT
1	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
2	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓
3	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
4	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
5	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
6	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
7	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
8	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
9	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
10	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
11	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
12	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
13	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
14	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
15	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
16	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗

First, the 15 units in which alliteration was found do not always include alliteration in the target units. In Bulgarian just 9 cases with alliteration are found among the observed units. The same is observed with Brazilian Portuguese, while in European Portuguese there are 8 units.

In these 15 units, the proper nouns are kept or adapted slightly (or transcribed in the case of Bulgarian) and the rendition of the definite descriptions into the lexicon of the target language is made by the literal or most common translation. Thus, in the cases where alliteration of the first

sound of the proper noun and the definite description occurs, it can be considered a coincidence. In no case is the proper noun of the character changed to create an alliteration with the existing noun or adjective in the lexicon of the corresponding language.

The situation with the cases of onomatopoeic alliteration and rhyme is similar. Rhyme appears in only one target unit in Brazilian Portuguese, in which, the source proper noun is copied and the head noun in the definite description is coincidentally has the same form as the source. Thus, the target unit includes the same words as in the source, with different word order.

With regard to the three cases of onomatopoeic alliteration in the source units only one is transferred to the target units. This happens in all three target units, although it can also be considered a coincidence to some extent because the noun form used to translate the source noun is an onomatopoeia on its own. The other cases of onomatopoeic alliteration are not transferred to the target units.

In summary, the translator is confronted with several limitations arising both from the nature of the referent of the units he is translating and from the specifications in the translation guidelines. Thus, one major limitation arises from the inherent characteristics and gender of the nouns denominating animals in Bulgarian and Portuguese, which prevent any change in the definitive description. In addition, the translator's room for manoeuvre is restricted by the rule that proper nouns may not be changed.

In scenarios where the translator finds a single noun in the definitive description, a simple approach is to add a characteristic feature to the character's name to achieve the phonetic effect present in the source. This is, of course, the case when the priority is on keeping the phonetic effects. For example, if the definitive description consists only of a head noun, the translator's options are relatively limited. Nevertheless, the translator can insert additional elements to maintain the intended phonetic effect, e.g. *Ricardo the Racoon – Ricardo o Guaxinim Risonho.*

Another possible strategy is to change the proper noun of the character. However, this option comes with potential challenges, especially in the context of global distribution of the games and client preferences. Clients often want linguistic consistency in different languages or insist on keeping the proper noun. While changing the character's name might serve the purpose of maintaining phonetic impact, it is often rulled out by the client's instructions.

Therefore, it is important to recognise that despite these possible solutions and taking into account the limitations mentioned above, the successful transfer of phonetic effects into the target language often depends on fortuitous circumstances. This is the case, for example, when the noun in the source and target languages share a common etymological root, which facilitates

the retention of the phonetic effect. In most cases, however, these effects are lost.

6 Conclusion

This master's thesis is based on the work developed during the internship at Unbabel. The company provided an engaging internship experience that allowed the interns to collaborate, understand in depth the systems and resources used for translation tasks in the company, and make suggestions for their improvement. This experience sparked my interest in learning more about the challenges of translating multilingual glossaries, with a special focus on translating multiword names of characters.

After the detailed insight regarding the company where the internship took place, provided in Chapter 2, namely with regard to the translation processes and workflow, a description of the various tasks carried out during the internship was presented. Some important conclusions emerged from this experience. Firstly, the way the internship was organised enhanced the interns' enthusiasm and ability to complete tasks effectively, and provided a thorough understanding of the quality assurance and translation tools, their purpose and functionality. Secondly, the collaborative environment encouraged the exchange of new ideas and the development of critical thinking skills. Most importantly in the case of the current thesis, the glossary curation task provided valuable insights into key aspects of glossary creation and translation. This included reflections on the structure of the glossary units, identification of patterns in the source units and observations on how these were handled in the target units. This also included some considerations about the information encoded in the glossary and the need for additional information, as discussed in this work.

In Chapter 3, we built on relevant theoretical background by looking at key concepts in machine translation, computer-aided translation and the creation and management of multilingual glossaries. This chapter provided an insight into the historical development and current state of the art of machine translation systems, the functions of computer-aided translation tools and various approaches to handling glossaries, as well as reflection on multiword proper nouns for special purposes. We have also highlighted relevant distinctions between glossaries designed for human use and those intended for machine use.

Chapter 4 described the methodology used for curating the glossary and selecting the glossary

units for analysis. The aim of the curation was to analyse the structure of the glossary and its units and to identify translation challenges. This chapter also covered some aspects of data anonymisation in accordance with data protection regulations. The steps of the process included data collection, description of some of the resources used, selection of the glossary units to be analysed and the criteria for doing so, and an initial analysis that included various aspects such as the structure of the units and some orthographic considerations. The selection of glossary units foresaw the analysis of units that could be more challenging for machine translation and human post-editing, namely multiword units. An additional level of difficulty was identified for a particular group of multiword units that represented names of characters appearing in the client's video games.

Chapter 5 formed the core of our study and provided an in-depth analysis of the selected glossary units. The morpho-syntactic structures of the source and target units were described, the translation strategies used were examined and deviations and inconsistencies with regard to orthography, punctuation, gender and word order were considered. Our analysis revealed that the morpho-syntactic structure of the source units consistently followed the pattern [PROPER NOUN] + [DEFINITE DESCRIPTION]. In contrast, the structure of the Bulgarian and Portuguese target units showed considerable variations. We compared structures of the source and target units and examined the translation strategies used. We also considered word order and semantic changes resulting from changes in word order.

Furthermore, we explored the presence of phonetic effects within glossary units and their importance, particularly in the context of video game character names. We found that phonetic effects such as alliteration, rhyme and onomatopoeic alliteration are present in more than half of the units analyzed (57%). These phonetic effects are not limited to literature, but also occur in slogans, wordplay, nursery rhymes and song lyrics. In our case, they occur in the names of video game characters, and they aim to increase the memorability of characters' names, make the games more engaging, and connect the characters to the unique universe of each game. The games considered in this work belong to the "casual and social games genre", and appeal to an audience that appreciates the playful and creative aspects of various game elements, including character names. Consequently, these names are created with playfulness in mind, and translators should strive to uncover the embedded meanings and effects in the source units and mirror them in the target units. However, it is often not possible to transfer all the nuances and facets of the source to the target unit.

In addition, this chapter highlighted the importance of context in the translation process, especially when assigning gender to multiword proper nouns and the considerable amount of

time and research that is often required to inform translation decisions regarding this aspect. Assigning gender to proper nouns is particularly challenging for translators when the source language (in our case it is English) does not mark nouns and adjectives according to gender, whereas they are marked in the target languages (Portuguese and Bulgarian). Translators have to intentionally assign gender to target units or reconcile unintentional gender differences (i.e. the inherent gender of certain nouns in Bulgarian and Portuguese with the perceived gender of conventional proper nouns). In this process, it was often necessary to research publicly available sources regarding the games to confirm or determine gender assignments for uncommon proper nouns. This highlighted the importance of context in translating video game names and the significant amount of time and effort translators need to spend on research, when the glossary does not provide sufficient context, as well as the potential errors that can occur if they do not have the time to research or access the information they need.

Chapter 5 also contained some observations regarding best practices on the creation and translation of multiword proper nouns in multilingual glossaries, from which we underline the following:

1. Recommendations for glossary creators:

- Associate enough context and a detailed description (including images when relevant) to source glossary units
- For languages that use different scripts (e.g. Latin and Cyrillic), give the translator the possibility of choosing between transliteration, transcription and translation of proper nouns, if the client agrees. For multiword proper nouns, indicate which elements should be transliterated, transcribed, translated or kept in their original form.

Glossary creators should provide translators with specific information that is crucial to their work. Ideally, this information should be present in the source glossary and should include elements such as a brief character description and an image of the character. Many of the challenges in understanding the character's semantic context arise from the fact that these exist in a fictional universe that is separate from the real world. For this reason, access to additional context is essential. Given that these characters only exist in the universe of the game in question, translators lack real-world knowledge about them and must rely on the in-game context and accompanying information to acquire this knowledge. Without access to this context and information, it can become extremely difficult to make informed translation decisions, a scenario that is bound to affect the quality of the translation produced.

In our case, some of the characters are anthropomorphic, so real-life knowledge can provide some context. However, there are also cases where the characters are entirely fictional. In such cases, translation decisions become very challenging, if not impossible, without any context. The reason for this challenge lies in the potential ambiguity or ambivalence of these characters names, which must be clarified (or maintained in some way) in the target terms.

In other contexts, accessing or referring to real-world knowledge may be straightforward or unnecessary because it is already part of general knowledge. In the particular case discussed in this work, however, we are dealing with entirely fictional characters and worlds. It is, thus, important to be able to access both an image and a description of the character. The extent to which these dimensions and effects are incorporated into the character's name or the definite description that underlines some of its characteristics can vary and require additional research and context. Consequently, the translator's goal is to avoid losing facets of the meaning of the source unit or the fantasy effects and elements that it incorporates, as mentioned below, along with some other recommendations for translators:

2. Recommendations for glossary translators and for localisation:

- Use punctuation marks and capital letters in a coherent way according to the grammar of the target language.
- Pay attention to the definite article used in the target units, especially when different forms are available.
- When translating multiword proper nouns in languages with grammatical gender, pay attention to the gender of each element and look for additional context (if not enough context is available in the glossary) in order to accurately translate it.
- When creative elements are identified in source units, consider all hidden meanings or effects and define priorities regarding which ones need to be retained when it is not possible to maintain all of them in the translation.
- Identify all units in the same glossary that have a similar structure or characteristics and verify if a common approach can be applied.

In summary, a significant finding presented in Chapter 5 was the prevalence of phonetic effects in the names of video game characters, their role as playful and creative elements of the source units and the need to maintain them in the translation. This highlighted the importance of prioritising and retaining embedded meanings and effects during translation. Furthermore, it

illuminated the complexities of translating from an unmarked language in terms of gender into gender-marked languages and emphasized the need for context-rich glossaries.

Given the scope of this thesis and the time constraints, several promising areas for further investigation remain uncharted. One these is the application of the conclusions and recommendations derived from the analysis of the 28 source units and their translations considered in this work to other sets of multiword glossary units containing, for example, a broader set of multiword proper nouns. This would allow for assessing the generality of the results presented here to a wider range of units.

Looking ahead, this thesis paves the way for future studies of multiword glossary units. It has also identified interesting topics in the use of glossary units by humans and machines, especially in the case of creative names of video game characters. Further research in this area promises to provide valuable insights into how machine translation deals with human creativity in the domain of video games.

In summary, this Master's thesis has deepened our understanding of multilingual glossary curation and translation challenges, while shedding light on the complex world of multiword names of video game characters. The insights, recommendations, and opportunities for future research presented here contribute to the ongoing discussion about multilingual glossary management and have the potential to improve translation quality in specialised fields like the translation of texts related to video games and video game localisation.

Appendix A

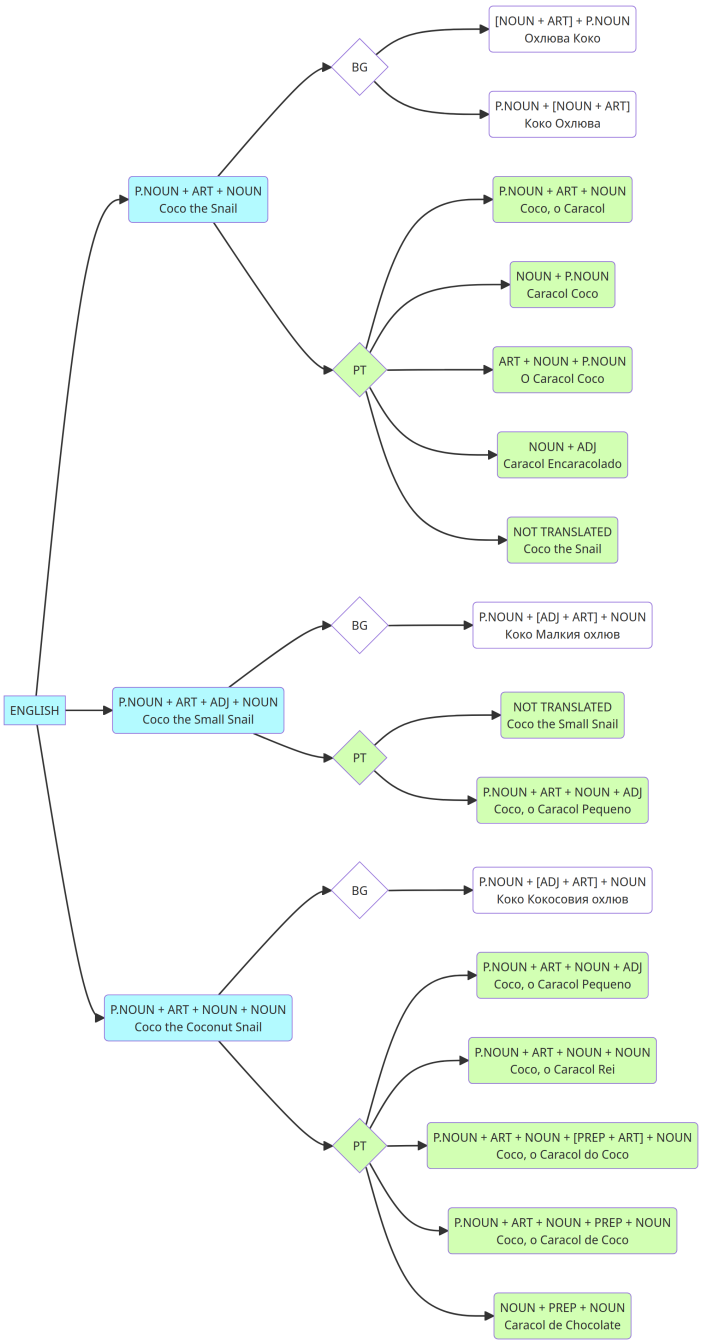

Figure A.1: Comparison of the structures of the glossary units and pseudo-examples

Bibliography

Automatic Language Processing Advisory Committee (ALPAC). (1966). *Language and Machines: Computers in Translation and Linguistics* (Tech. Rep.). Washington, D. C.: National Academy of Sciences.

Baldwin, T., & Kim, S. N. (2010). Multiword Expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (pp. 267–292). Boca Raton: CRC Press.

Bar-Hillel, Y. (1960). The Present Status of Automatic Translation of Languages. *Advances in Computers*, *1*(C), 91–163. Retrieved from `http://www.mt-archive.info/Bar-Hillel-1960.pdf`

Bar-Hillel, Y. (1971). Some Reflections On The Present Outlook For High-Quality Machine Translation. *Feasibility Study on Fully Automatic High Quality Translation*, 73–76.

Bernal-Merino, M. (2014). *Translation and Localisation in Video Games* (Vol. 6).

Bertills, Y. (2003). *Beyond Identification: Proper Names in Children's Literature* (Unpublished doctoral dissertation).

Bowker, L. (2002). *Computer-Aided Translation Technology: A Practical Introduction.* Ottawa, Canada: University of Ottawa Press.

Bowker, L., & Ciro, J. B. (2019). *Machine Translation and Global Research.* Emerald Publishing Limited.

Bussmann, H. (1996). *Routledge Dictionary of Language and Linguistics.* London/New York: Routledge.

Buzan, T., & Buzan, B. (2006). *The Mind Map Book.* BBC Active.

Cabré, M. T. (2010). Terminology and translation. In Y. Gambier & L. Van Doorslaer (Eds.), *Handbook of translation studies vol. 1.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Cabré, M. T., & Sager, J. C. (1999). *Terminology: Theory, Methods, and Applications.* Amsterdam / Philadelphia: J. Benjamins Publishing Company.

Chatfield, T. (2010). *Fun Inc.: Why Gaming Will Dominate the Twenty-First-Century.* New York: Pegasus Books.

Collection of Electronic Resources in Translation Technologies. (2012). *Glossary of translation tool types.* Retrieved from `http://aix1.uottawa.ca/{~}certt/CERTT-main-EN.htm`

Comparin, L., & Mendes, S. (2017a). Error detection and error correction for improving quality in machine translation and human post-editing. In *Proceedings of the 18th international conference on intelligent text processing and computational linguistics – cicling 2017.* Springer Publishing Company. Retrieved from http://hdl.handle.net/10451/33007

Comparin, L., & Mendes, S. (2017b). Using error annotation to evaluate machine translation and human post-editing in a business environment. Retrieved from https://ufal.mff.cuni.cz/eamt2017/user-project-product-papers/papers/user/EAMT2017{_}paper{_}76.pdf

Cunha, C., & Cintra, L. (2017). *Nova Gramática do Português Contemporâneo* (P. Greiger, Ed.). Rio de Janeiro: Lexicon.

DeepL. (2022a). *About the glossary feature.* Retrieved 2022-11-11, from https://support.deepl.com/hc/en-us/articles/360021634540-About-the-glossary-feature

DeepL. (2022b). *DeepL Subscription Plans.* Retrieved 2022-11-11, from https://www.deepl.com/pro?cta=menu-plans/

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI). (2014). *Multidimensional Quality Metrics Definition.*

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., & Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (Vol. 1: Long Pa, pp. 1370–1380). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P14-1129

Dicionário Prático Ilustrado. (1959). Porto: Lello e Irmãos Editores.

DuPont, Q. (2018). *The Cryptological Origins of Machine Translation.* Retrieved 19/06/2019, from http://amodern.net/article/cryptological-origins-machine-translation/{\#}rf51-10627

European Parliament, & Council of the European Union. (2016). *General Data Protection Regulation (GDPR)* (Vol. 59). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:119:FULL

Faculdade de Letras da Universidade de Lisboa. (2018). *Normas regulamentares do Mestrado em Tradução.* Retrieved 2019-01-31, from https://www.letras.ulisboa.pt/pt/documentos/cursos/mestrados/3081--1092/file

Fernandes, L. (2006). Translation of Names in Children's Fantasy Literature: Bringing the Young Reader into Play. *New Voices in Translation Studies*, *2*, 44–57.

Hasler, E., De Gispert, A., Iglesias, G., & Byrne, B. (2018). Neural Machine Translation Decoding

with Terminology Constraints. In *The 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies.* New Orleans, Louisiana: The Association for Computational Linguistics. Retrieved from `http://arxiv.org/abs/1805.03750`

Hermans, T. (1988). On Translating Proper Names, with reference to De Witte and Max Havelaar. In M. Wintle (Ed.), *Modern dutch studies: Essays in honour of professor peter king on the occasion of his retirement* (pp. 11–24). London: Bloomsbury Academic.

Hutchins, W. (1986). *Machine translation: past, present, future.* New York: Halsted Press.

Jaekel, G. (2000). Terminology Management at Ericsson. In R. C. Sprung & S. Jaroniec (Eds.), *Translating into success: Cutting-edge strategies for going multilingual in a global age* (Vol. XI, pp. 159 – 171). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Koehn, P. (2017). Neural Machine Translation. In *Statistical machine translation* (p. 117). Retrieved from `http://arxiv.org/abs/1709.07809`

Kvillerud, R. (1985). Personnamnens betydelse, form och funktion hos författaren Maria Gripe. In G. Hallberg, S. Isaksson, & B. Pamp (Eds.), *Norna-rapporten 34. nionde nordiska namnforskarkongressen. lund 4-8- augusti 1985.* (pp. 51–62). Uppsala: NORNA- förlaget.

Laporte, É. (2018). Choosing Features for Classifying Multiword Expressions. In M. Sailer & S. Markantonatou (Eds.), *Multiword expressions: Insights from a multi-lingual perspective* (pp. 143–186). Berlin: Language Science Press,.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica tecnologies de la traducció*(12), 455–463. doi: 10.5565/rev/tradumatica.77

Machado, J. P. (2003). *Dicionário Etimológico da Língua Portuguesa.* Lisboa: Livros Horizonte.

Maclean, M., Bryant, P., & Bradley, L. (1987). Rhymes, Nursery Rhymes and Reading in Early Childhood. *Nerrill-Palmer Quarterly*, *33*(3), 255–281.

Makkai, A. (1972). *Idiom Structure in English.* The Hague, The Netherlands: Mouton.

Masini, F. (2019). *Multi-Word Expressions and Morphology.* Oxford Research Encyclopedias, Linguistics. doi: 10.1093/acrefore/9780199384655.013.611

McGillis, R. (1996). *The nimble reader: literary theory and children's literature.* New York: Twayne Publishers.

Melby, A. K. (2012). Terminology in the age of multilingual corpora. *Journal of Specialised Translation*(18), 7–29.

Mendes, A., & Antunes, S. (2016). Collocations in Portuguese: A corpus-based approach to

lexical patterns. *Collocations cross-linguistically*, 141–166.

Merriam-Webster Dictionary. (2019). Retrieved 25/11/2019, from `https://www.merriam-webster.com/`

Milne, A. A. (1926). *Winnie-the-Pooh.* London: Methuen & Co. Ltd.

Milne, A. A. (1987). *Mecho Puh (Translated by Vera Slavova).* Sofia: Otechestvo.

Newmark, P. (1988). *A Textbook of Translation.* Shanghai Foreign Language Education Press.

Nord, C. (2003). Proper names in translations for children: Alice in Wonderland as a case in point. *Meta*, *48*(1-2), 182–196.

O'Hagan, M., & Mangiron, C. (2013). *Game Localization: Translating for the global digital entertainment industry.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Pearce, M. (2007). *The Routledge Dictionary of English Language Studies.* London and New York: Routledge.

Poddar, L. (2013). *Multilingual Multiword Expressions* (Unpublished doctoral dissertation). Indian Institute of Technology Bombay, Mumbai.

Poesio, M., & Vieira, R. (1997). A Corpus-Based Investigation of Definite Description Use. Retrieved from `https://arxiv.org/pdf/cmp-lg/9710007.pdf`

QT21 Consortium. (2018). *QT21 – Quality Translation 21.* Retrieved 2018-11-23, from `http://www.qt21.eu/?target=Introduction`

Reinke, U. (2018). State of the art in Translation Memory Technology. In *Language technologies for a multilingual europe tc3 iii.*

Rogers, M. (2006). *Terminology, Term Banks and Termbases for Translation* (2nd ed.). Elsevier Science. doi: 10.1016/B0-08-044854-2/00477-6

Roturier, J. (2019). XML for translation technology. In M. O'Hagan (Ed.), *The routledge handbook of translation and technology.*

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 1 – 15). Mexico City: Springer.

Shiffman, D. (2012). Neural Networks. In *The nature of code: Simulating natural systems with processing* (p. 520). The Nature of Code. Retrieved from `https://natureofcode.com/book/chapter-10-neural-networks/`

Smadja, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*, *19*(1), 143–177.

Sprung, R. C., & Jaroniec, S. (Eds.). (2000). *Translating Into Success: Cutting-edge strategies for going multilingual in a global age.* Amsterdam / Philadelphia: John Benjamins Publishing

Company.

Stein, D. (2018). Machine translation: Past, present and future. In G. Rehm, F. Sasaki, D. Stein, & A. Witt (Eds.), *Language technologies for a multilingual europe: Tc3 iii* (pp. 3–17). Berlin: Language Science Press. Retrieved from http://langsci-press.org/catalog/book/106

Stoyanov, S. (1983a). Prilagatelno ime [Adjective]. In *Gramatika na savremennia balgarski knizhoven ezik. tom ii morfologia [grammar of the contemporary standard bulgarian language. volume ii morphology]* (pp. 146–178). Sofia: Bulgarian Academy of Sciences.

Stoyanov, S. (1983b). Sashtestvitelno ime [Noun]. In *Gramatika na savremennia balgarski knizhoven ezik. tom ii morfologia [grammar of the contemporary standard bulgarian language. volume ii morphology]* (pp. 42–146). Sofia: Bulgarian Academy of Sciences.

Stoykova, V. (2011). Interpreting Bulgarian sound alternations of inflectional morphology in DATR. *ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, *1*, 486–491.

Unbabel. (2017a). *A closer look at Unbabel's award-winning translation quality estimation systems.* Retrieved 2018-11-19, from https://unbabel.com/blog/unbabel-translation-quality-systems/

Unbabel. (2017b). *Language Guidelines.* Retrieved 2019-02-04, from https://help.unbabel.com/hc/en-us/sections/205912528-Language-Guidelines

Unbabel. (2018a). *Evaluation System.* Retrieved 2018-10-09, from https://help.unbabel.com/hc/en-us/articles/217608238-Evaluation-System-FAQ

Unbabel. (2018b). *Smartcheck, Glossaries and Translation Memories.* Retrieved 2018-09-16, from https://help.unbabel.com/hc/en-us/articles/360003342214-Smartcheck-Glossaries-Translation-Memories-and-Quality-Estimation

Unbabel. (2018c). *Unbabel developers documentation.* Retrieved 2018-06-01, from https://developers.unbabel.com/v2/docs

Unbabel. (2018d). *What are Glossaries?* Retrieved 2019-02-01, from https://help.unbabel.com/hc/en-us/articles/360003366273-What-are-Glossaries-

Unbabel. (2018e). *What is an Editor?* Retrieved 2018-09-12, from https://help.unbabel.com/hc/en-us/articles/360003342914-What-is-an-Editor-

Unbabel. (2019). *Unbabel - Translators.* Retrieved 2019-01-31, from https://unbabel.com/translators/

Universal Dependencies, & Nivre, J. (2017). *Universal POS tags.* Retrieved 2018-10-14, from http://universaldependencies.org/u/pos/index.html

Weaver, W. (1955). Translation. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages: fourteen essays* (pp. 15–23). New York: Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass., and John Wiley & Sons, Inc.