Universidade de Lisboa
Faculdade de Letras

Analysis of context-aware Translation Memories:
Part-of-Speech pattern distribution and gender neutral Translation Memories

MARJOLENE HAIDÉ MACHADO PAULO

Relatório de estágio especialmente elaborado para a obtenção do grau de Mestre em Linguística, orientado pela professora Doutora Helena Gorete Silva Moniz e coorientado pela Doutora Vera Mónica dos Santos Cabarrão

2022

Dedicatória

Para a minha mãe,
o meu maior exemplo
de esforço e dedicação.

Acknowledgments

Às minhas orientadoras, Professora Doutora Helena Moniz e Doutora Vera Cabarrão pela orientação, pelo apoio e pela confiança que depositaram em mim ao longo do estágio. Por todo o valioso *feedback* que contribuiu para o meu crescimento. Por despertarem em mim o interesse na área da Tradução Automática.

À Unbabel, especialmente à equipa de NLP, pela amabilidade e disponibilidade que me demonstraram ao longo destes meses.

Aos meus amigos, que acompanharam de perto esta jornada. Por terem aturado os meus momentos menos bons, mas que sempre me deram a força necessária para continuar.

À minha família, por todo o amor e apoio incondicionais. Sou-vos muito grata por terem acreditado em mim.

Table of Contents

Abstract

Translation Memory is the most commonly used Computer-aided Translation system, whose main purpose is to store and retrieve previous high-quality translated sentences. Although they are very important systems for the translation process, they operate at the sentence level. This approach can be problematic as translated documents without considering the full context can cause coherence and cohesion issues at the text level.

The present thesis aims at the analysis and creation of context-independent and gender neutral Translation Memories by resorting to part-of-speech (POS) information, in order to automatically identify context-dependent segments. This would enable the reuse of segments without causing meaning constraints at the document level. In order to achieve this purpose, three experiments were conducted focusing on customer support data.

We firstly conducted a pilot experiment to annotate context-related issues in a dataset of 2,045 TMs for Brazilian Portuguese and European Spanish. The results showed that gender agreement was the most frequent category (80%), followed by Register (20%).

For the second experiment, we analyzed a total of 5,200 segments for Portuguese, Brazilian Portuguese, Spanish and Latin-american Spanish, with English as the source language. The goal of this analysis was to annotate context related issues with a new context annotation typology. Thereafter, all the context-dependent data was analyzed by a POS tagger in order to understand if it was possible to create sequences of parts of speech patterns that could distinguish context-dependent TMs from context-independent ones.

The last experiment consisted on the analysis of a new dataset of a total of 8,000 segments for the same languages. The goal of this experiment was to verify if the previously found patterns could actually identify context-dependent segments.

Results showed that 1,298 out of 15,245 TMs were context-dependent in which 1,263 had gender constraints. We were able to turn the latter segments into gender neutral, therefore improving 8% of very frequent data.

Keywords: Translation Memory; context; parts of speech; customer support domain

Resumo

As Memórias de Tradução são as ferramentas de Tradução Assistida por Computador mais comuns, cuja funcionalidade é o armazenamento de pares de frases e as suas respectivas traduções, permitindo que sejam recuperadas a qualquer instante durante o processo de tradução. Embora sejam importantes para a organização e gestão do texto, uma desvantagem destes sistemas prende-se com o facto de operarem ao nível da frase. Esta abordagem pode ser problemática, uma vez que a tradução de segmentos independentes, sem a consideração de relações de dependência intra e interfrásicas, pode causar problemas de coerência e coesão no texto.

O presente trabalho centralizou-se no módulo *TM server* da Unbabel, uma empresa *startup* portuguesa que fornece serviços de tradução recorrendo a Tradução Automática e outros sistemas de Inteligência Artificial, juntamente com uma comunidade global de tradutores que auxiliam no processo de pós-edição. O *TM server* é uma ferramenta muito importante para o processo de tradução, no entanto, por armazenar segmentos muito repetitivos e isolados de contexto, pode conter problemas que apenas podem ser resolvidos com mais informação para além da contida na frase. Por conseguinte, a presente tese visa a análise e criação de Memórias de Tradução independentes de contexto e neutras quanto ao género, recorrendo a informação de classificadores morfossintáticos, de forma a possibilitar a identificação automática de segmentos dependentes de contexto. Tal possibilitaria a reutilização destes segmentos sem causar qualquer tipo de constrangimento semântico ao nível do documento. Para tal finalidade, foram realizadas três experiências com base em dados de apoio ao cliente.

Primeiramente, foi realizado um estudo piloto, cujo corpus compreendia um total de 2045 segmentos com dados para o português brasileiro e para o espanhol europeu (ou peninsular), sendo o inglês a língua de origem. Com esta análise, pretendia-se a anotação de questões relacionadas com o contexto. Os resultados mostraram que a concordância de género foi a categoria predominante, representando 80% destes casos. Registo foi também uma categoria anotada, sendo atribuída aos restantes 20% dos dados. Esta análise permitiu a construção de uma tipologia de anotação de problemas relacionados com contexto com cinco categorias distintas, nomeadamente concordância de género, concordância em número, elipse, terminologia e registo.

A segunda experiência ocorreu em duas partes distintas. Na primeira parte, foram analisados um total de 5200 segmentos com dados para o português europeu (PT) e brasileiro (PT-BR), espanhol europeu (ES) e da América Latina (ES-LATAM). À semelhança do estudo piloto, esta fase consistiu na identificação e anotação de problemas relacionados com contexto, mas agora usando a nova tipologia de anotação. Como resultado, recolhemos um total de 338 segmentos dependentes de contexto. No que diz respeito a categorias de contexto, os resultados foram consistentes com os obtidos previamente, pelo que problemas relacionados com concordância de género continuaram a ser a maioria, correspondendo a 98% dos casos. Ao contrário dos resultados obtidos anteriormente, a categoria Registo apenas registou 1,2% dos casos. Em adição, foram também identificadas novas categorias, sendo estas Elipse e Terminologia que foram menos representativas do que as anteriores.

Na segunda parte desta mesma experiência, usando um classificador automático, foram analisados todos os segmentos dependentes de contexto, de forma a que fosse possível verificar padrões morfológicos que pudessem evidenciar dependência de contexto. Através da informação obtida pelo classificador automático morfossintático, foi possível a identificação de categorias gramaticais frequentemente envolvidas em problemas contextuais, sendo estes pronomes, adjetivos e verbos. Primariamente, quanto aos pronomes, os pronomes pessoais de terceira pessoa *-lo* e *-la*, para as variantes do português, e o pronome de primeira pessoa do plural *nosotros*, para as variantes do espanhol, foram bastante frequentes entre os dados. Quanto ao adjetivos, *satisfeito(a)*, *interessado(a)*, para PT e PT-BR, e *encantado(a)* e *emocionado(a)*, assim como outros adjetivos que permitem expressar agrado ou desagrado, para o ES e ES-LATAM, foram igualmente frequentes. Por último, a categoria gramatical "Verbo" foi muito frequente para o PT como para o PT-BR, correspondendo à ocorrência da expressão *Obrigado(a)*. Como tal, estas categorias gramaticais permitiram a criação de oito padrões, isto é, sequências de categorias POS, que permitem a identificação de sequências dependentes de contexto: três deles que ocorriam exclusivamente para PT e PT-BR e um exclusivo do ES e do ES-LATAM, sendo que os restantes mostraram serem comuns entre todas as variantes.

A fim de validar os padrões encontrados, foi conduzida uma terceira e última experiência, cujo objetivo era verificar se estes permitiriam a identificação de segmentos dependentes de contexto, numa amostra mais alargada. Para tal, esta análise compreendeu um total de 8000 TMs de sete clientes diferentes e de diferentes domínios de apoio ao cliente (i.e. gaming, tecnologia)

que nunca tinham sido analisados antes. Os resultados mostraram que, no total, dois dos oitos padrões permitiram a identificação de segmentos dependentes de contexto, sendo que um corresponde ao PT e PT-BR e um ao ES e ES-LATAM. Esta experiência permitiu concluir que três dos oito padrões permitem, de facto, a identificação de segmentos dependentes de contexto.

O presente projeto teve contribuições positivas. Por um lado, os resultados obtidos das tarefas de anotação permitiram o desenvolvimento de uma tipologia de anotação de contexto, desenvolvida com o intuito de auxiliar a comunidade de editores que trabalha diretamente com estes segmentos. Esta foi validada por profissionais na área que geram a documentação linguística na empresa e será implementada em breve.

Por outro lado, após uma análise de diferentes dados de diferentes domínios, foi notório que questões relacionadas com concordância de género foram as mais comuns entre todas as experiências. Por serem prevalentes, viu-se a necessidade de tornar neutros estes segmentos. Assim sendo, como tarefa final, foram apresentadas sugestões de traduções alternativas e neutras, mantendo sempre o sentido do texto original. A produção de segmentos neutros e independentes de contexto permite que estes sejam selecionados para fazer parte de qualquer documento, sem comprometer o significado global de um texto e, mais importante, sem causar dependências de género.

No global, foram analisados 15245 dados, sendo que 1298 destes eram dependentes de contexto. Dentre estes últimos, 1263 apresentaram questões relacionadas com género. Tornar neutros todas estas TMs, muito frequentes, permitiu reduzir em 8% estes casos.

Palavras-chave: Memórias de Tradução (in)dependentes do contexto, tipologias de estruturas associadas a contexto, classificador morfológico e distribuição de padrões, criação de memórias de tradução independentes do contexto, domínio de apoio ao cliente.

1. Introduction

Natural Language Processing (NLP) is a multidisciplinary field that converges knowledge from areas such as Linguistics, Artificial Intelligence and Computer Science to investigate how computers can process natural languages (Chowdhury, 2003). One of the most noticeable applications is Machine Translation. This approach began to be developed during the 1950s and has seen growth ever since. Some time after, during the 1990s, software powered systems that allowed the organization and management of previous translations were introduced commercially. These were called Computer-aided Tools. These systems have different tools integrated but one of the widely used ones is the Translation Memory.

In the context of industry, these segments have been used as complementary modules, in order to optimize the translation process, allowing cost reductions and enabling quick translations. However, a downside of these systems is that they have a sentence as a basic unit. This is mainly due to the fact that the system usually achieves human parity, that is, the output produced by the machine is close to one done by a translator (Hassan et al., 2018). However, when the sentences are put back together it often generates cohesion and coherence problems at a document level. This is problematic, as a text is a coherent and cohesive unit whose ideas are presented in a logical way.

Recent work done in this field aims to achieve a context-aware MT system by incorporating more context than the current sentence by tackling discourse phenomena in a document, and overall, reduce the problem that results from lack of contextual information. For the current work, we are not going to test an MT system itself, rather we intend to analyze segments that have been previously translated and therefore stored in the database, namely the *TM server*, in order to understand which categories may generate context problems.

For the scope of this dissertation, our assumption is that by using part-of-speech information it would be possible to automatically identify context-dependent segments, therefore, simplifying the overall process of curation and incrementing the productivity of the post-edition process. Also, we wanted to turn very frequent segments with gender constraints into gender neutral Translation Memory, in order to re-use these segments as many times as possible without meaning constraints. The work was done at Unbabel and was centered around customer support domains through emails.

The thesis is organized in the following way: in chapter 2, we provide a description of the host company, Unbabel; in chapter 3, we present a brief historical overview of Machine Translation and Translation Memory systems, as well as the current work that is being done to achieve document-level and context-aware translation; in chapter 4, we describe the methodology used for the experiments conducted; in chapter 5, we present the results obtained for all of the experiments; in chapter 6, we propose some suggestions on turning the TMs with gender constraints into gender neutral TMs; and, lastly, chapter 7 shows the conclusions from the analysis and intents for further work.

2. Unbabel

Unbabel is a company founded in 2013 by Vasco Pedro, João Graça, Hugo Silva, Sofia Pessanha and Bruno Prezado. It is headquartered in Lisbon and has several offices scattered in various countries, such as the United States, United Kingdom, Romania, and Philippines. The company's main goal is to provide translation services mainly focused on customer support by combining artificial intelligent (AI) systems with the help of translators to deliver rapid and high-quality translations to their customers, and, overall, to make the translation process more efficient. In a sense, Unbabel is a route that connects a company and its customers, eliminating the communication barriers generated by language differences.

In December 2021, Unbabel acquired Lingo24, a company that provides translations and other language related services across domains, such as marketing, travel and press releases. It was founded in 2001, in Edinburgh, and also worked with a large community of translators and linguists who are proficient or native speakers from a particular language pair. At the moment, both companies are working towards a unified business model in order to provide multilingual content at a global scale for a broad set of domains.

In the following sections, it will be presented in more detail the content types translated, the translation processes, as well as the quality control processes used in the company.

2.1. Company's workflow

The content types translated at Unbabel are tickets (support emails between agents and clients), chat (instant messages), and frequently asked questions (FAQ), a list of commonly asked questions that can be found on a customer's website and are often available in different languages. Although the translation process is generally very similar to all content types, still there are specific particularities for each content type, such as different delivery times and quality requests. The translation process for every content type begins after a client's order has been submitted, either through a Customer Relationship Management (CRM) platform (e.g.: Zendesk, Salesforce) or through a platform provided by Unbabel. These platforms help manage all the interactions that occur between a customer support agent and their clients. In the next sections, we will describe in more detail the workflow of Unababel's translation processes.

2.1.1. Translation pipeline

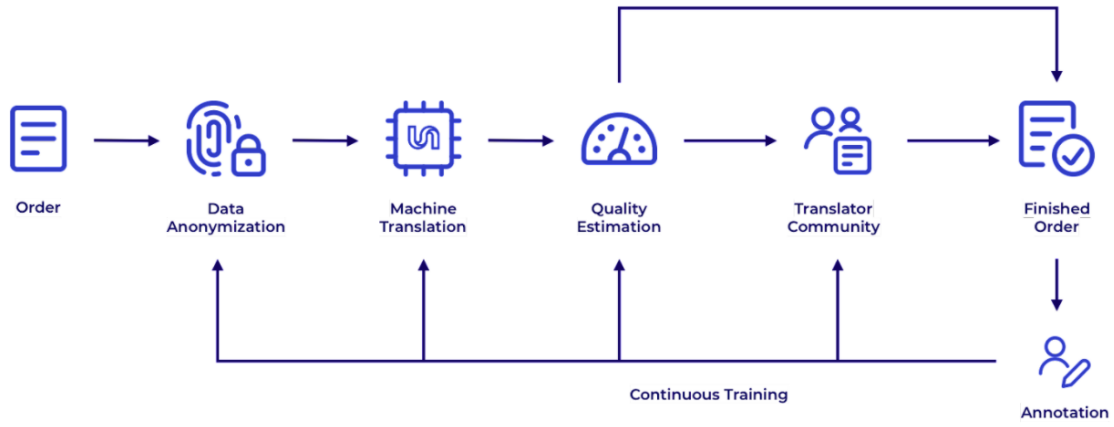

Figure 1 *- Unbabel's Translation pipeline for tickets*

Figure 1 illustrates the translation pipeline used at Unbabel for tickets, even though certain steps can be applicable to all content-types. After an order has been placed, there is an important step that consists of the document preparation. In this stage, the source text is analyzed in order to detect the language to be translated, as well as converted into a format that can be understood by the MT system (i.e. TXT, HTML and XLIF). The text is then split into sentences, as the translation is done at the sentence level. The following step consists in data anonymization. To ensure the client's privacy and in order to follow the General Data Protection Regulations (GDPR)[1], all Personally Identifiable Information (PII), such as proper names, emails, phone number or passwords, are identified by a Named Entity Recognition (NER) system and temporarily replaced by a placeholder. These can either be a generic placeholder, such as [EMAIL] or [PHONE NUMBER], or, in the case of proper names, a semantic equivalent, a generic proper name or family name that agrees in gender and case with the original name. Once the anonymization is completed, it is verified all the client's information, such as register and specific terminology to be used in the translation, as well as Translation Memories (TMs) collected in a server, a database that stores sentences that are frequently translated or that have been curated by terminologists, expert linguists or translators that ensure high quality translations. This tool is useful for retrieving and reusing previous translations, thus

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

saving time and costs in the translation process. After all these processes have been applied, the text is then translated by a Neural Machine Translation system (NMT), which is trained to adapt to Unbabel's clients (Carrozo, 2017).

After the full document is translated, it is sent to the Quality Estimation (QE) module. This is also a machine learning software that sets a score for the quality of the translation. If the translation is above a very conservative threshold, then it is considered as a QE skip and it is sent directly to the customer. Otherwise, if the QE score is following the stipulated threshold, the translation is sent to the community to be edited. Only after that, the translation, now reviewed, is sent to the customer.

Unbabel works with a global community of professionals and non-professional editors. The non-professional editors are bilinguals, who are proficient in a particular language pair and who assist with post-editing the outputs generated by the MT. They edit and improve the translations and adjust them to the client. The professional community is formed by terminologists and senior editors who are professional linguists and or professional translators and who are in charge of producing and reviewing linguistic resources for the company. Their work includes the curation of TMs, a process that involves the correction of errors in the segment so they can be reused. Having a large community such as this allows for different editors to work on different projects at the same time, thus reducing the delivery times.

The final step in the pipeline is the annotation process. This process consists of identifying and labeling errors according to a specific typology, and is extremely important as it allows the retraining of the artificial intelligence (AI) systems, namely the MT, NER and QE, and assures the quality of the pipeline (this process will be presented in more detail in section 2.2.2).

The translation pipeline for chat messages is different from the ticket one, in the sense that less steps are required, since it heavily relies on specific pretrained models for chat based on post-edited data. This content-type demands a rapid response, consequently, the translated text is not sent to an editor nor QE is applied. After MT, the translation is sent directly to the customer. Nevertheless, a sample of the translation is sent to annotation for errors' evaluation and all the edited data is fed to retrain the MT models so they can improve.

Finally, FAQs is a content-type used to improve customer experience, as it acts as a self-service option to clients seeking solutions proactively, thus it requires a high translation

quality as website content and the public face of the client. As demonstrated in *Figure 2*, after the document is translated by the MT, it is divided into sections and sent individually to the editors' community to be edited. Here, an extra step is added to the pipeline after the Post-edition (PE). Once translated and edited, the full document is sent to a senior editor for proofreading to make sure that the text is consistent and cohesive. After this process, the document is sent to the customer.

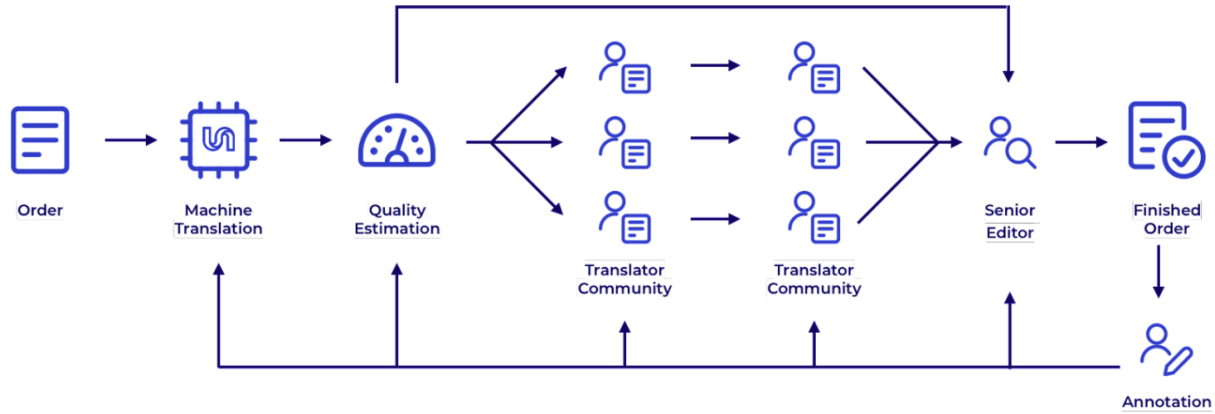

Figure 2 *- FAQs translation pipeline*

At the moment, Lingo24 content is running in FAQ's pipeline since there are new content-types, thus ensuring the best quality for the translations.

2.1.2. Annotation and evaluation process

For all content-types, after the order has been submitted to the customer, a sample of the translation is annotated by Unbabel's community of annotators, built by professional linguists and translators. The annotation process consists of the classification of the errors according to a specific typology and is extremely important as it allows to understand the quality and performance not only of the MT, but also of the community of editors, since both MT and PE are annotated. Subsequently, all the annotated data is used to retrain the MT, QE, and NER systems, thus allowing for continuous improvements.

For the error classification, the company uses an adaptation of the Multidimensional Quality Metrics[2] (MQM). MQM is a framework commonly used in the industry for quality assessment and allows us to understand the types of errors that occur in the translation and their severity (Lommel et al., 2014). In this typology, the errors are categorized according to three major categories: Accuracy, Fluency and Style. Firstly, *Accuracy* is related to how accurately the target text reflects the source text. *Fluency* refers to how natural a translated text sounds. Lastly, *Style* is related to stylistic problems found in the text. Also, to each error a severity is assigned: Neutral, Minor, Major and Critical. Neutral is a severity assigned to the source text, whenever there is an error in it. Minor errors are usually related to punctuation or orthography, hence they do not compromise the message of the text. Errors classified as major are more serious than the previous. To some degree, they affect the original meaning of a text. Finally, critical errors are severe in the sense that they not only compromise the original text but also mislead the reader and may even lead to security or legal implications (Sanchez, 2020).

Another important step for quality control is the evaluation of the editors' community. Periodically, quality audits are conducted to measure the performance of the editors. This evaluation is performed by professional translators and linguists who comment and rate the editors' translations from 1 (a bad quality translation) to 5 (an excellent translation), as illustrated in *Figure 3*:

Figure 3 *- Editor's assessment*

This process is conducted to ensure that the editors are producing good quality translations. When joining the company, all editors start with a testing phase where they get

[2] https://www.w3.org/community/mqmcg/2018/10/04/draft-2018-10-04/

familiar with the tools used and also the language guidelines in which they will work. After this stage, they move to a training phase. If their performance is positive, they move to paid content. On the contrary, if their performance is poor and they do not meet the criteria required for the company, they remain in this phase. It is important to mention that at all stages they are given feedback on their task performance in order to improve. This is also an extremely important step as it assures that the community is submitting the highest quality content possible.

2.2. Natural Language Processing team

The present internship allowed work alongside the Natural Language Processing (NLP) team. This is a multidisciplinary team formed by engineers, artificial intelligence experts and linguists who are in charge of all the processes that precede and succeed the MT step, called the *build* and *rebuild* stages. The first corresponds to the step prior to MT, where the document is pre-processed and prepared to be translated by the MT engine, as detailed on section 2.1.1. In contrast, the rebuild stage occurs after the MT, where the translated document is restored to its original form before it can be sent to the customer.

This team is also in charge of maintaining the necessary tools for the pre and post-processing of the texts, such as the already mentioned NER system, as well as, for example, the modules that allow sentence splitting, tokenization, and word alignments. Another important tool that the NLP team owns and that was the focus of the present internship is the Translation Memories (TM) server. A detailed description of this tool can be found in the following section.

2.2.1. Translation Memory Server

The present work focused on the TM server, a tool used in the build stage, the preprocessing stage prior to MT. This is a dynamic database that stores pairs of texts and their high quality translations, usually sentences (source and target), and allows re-using them as needed. The system by itself does not have the ability to translate a text, yet it is used in conjunction with the MT.

The retrieval of a segment is done by automatically comparing a source text with the data stored in the database and matching it with the most accurate target text. Various thresholds,

corresponding to the percentage of similarity between a source text and the respective TM entry, can be set for the retrieval of segments. There are two types of matches frequently used at Unbabel: exact match and fuzzy match. An exact match occurs when a source text matches exactly one entry stored in the database. In contrast, the fuzzy match occurs when the text has a certain similarity to one found in the database, however not fully. To obtain a fuzzy match, the first step is to remove the punctuation and lowercase the text. Also, to facilitate the retrieval, all TMs have relevant metadata associated with them, such as the client identification, the brand name, register and the content-type. The server then returns all the TM entries that are compatible, and provides the corresponding high quality translation.

In the company, there is also another specific feature used to retrieve TMs called "placeholder feature". As aforesaid, placeholders make it possible to temporarily replace sensitive information with either a generic word or by a semantic equivalent in the case of proper names. This can be seen in the following example:

(1) Source TM: Please have the 10€ by 11/03/2021.
Placeholder TM: Please have the CURRENCY-0 by DATE-0.

TM match:
Source TM: Please have the 20€ by 20/10/2025.
Source TM: Please have the $10 by 10/10/2010.

As shown in the example, when searching for a TM entry with any variable information, such as a monetary value or a date as in (1), the system automatically replaces it with a placeholder. The server then returns several other TM entries that resemble the source text. All of these steps take place before the MT and is very important as it allows to improve the TM matching rate.

2.3. Project goal

TMs are an important component in the company's translation processes. Since they are very repetitive content, they allow for the reduction of the delivery time of the translations,

guaranteeing quality and consistency throughout the repeated content. After the translation of a document, TMs are stored in the TM server in segments (which may not map directly to a sentence), meaning that they are isolated from context. This may imply that they may contain issues that can only be solved with information that precedes or succeeds them.

The goal of the internship project was to create gender neutral and context-aware segments by using not only POS information, but also context information (at the document level), in order to ensure that they could be freely selected to be part of any document without compromising the overall meaning of the text and without gender constrictions, thus meaning that they could be used in any contexts.

It is important to mention that, in the company, curated TMs may or may not be blocked for edition. A segment blocked for edition cannot be modified by the editors. Among other factors, this is a way to ensure that unnecessary changes that could possibly compromise the quality of the segment are made. For instance, the TM in (1) is completely correct, thus no further changes are needed. This means that this segment can be blocked from edition. On the contrary, an unblocked TM can be edited by an editor at any time and whenever needed. For instance, in (2) gender information is required in order to provide a correct and context-aware translation, hence, this segment needs to be unblocked so that the editors can perform the necessary changes to it.

(1) Source: But don't worry, they'll get to it very soon!
Target: Mas não se preocupe, isso vai ser resolvido em breve!

(2) Source: We will try to help you.
Target: Vamos tentar ajudá-**lo**.

3. State of the art

Machine translation is an important task in natural language processing that has seen exponential growth in recent years. MT systems are very powerful, constantly expanding in the market and can be seen as an aid to professional translators. In the first section of this chapter, we will present a brief history of MT and the types of approaches used during decades (section 3.1). Subsequently, we will describe the most widely used computer-aided translation tools, the Translation Memory (section 3.2.), as well as its history (3.2.1), common types (3.2.2.) and their potential limitations (3.2.3). Also, a brief discussion of the connection between MT and TM systems will be conducted (section 3.3) and how they are related to context. The final two subsections will tackle a definition of context (3.3.1) and the typologies for annotating context-related issues (3.3.2). In the last section (3.4), we present an also common NLP application namely, the Parts-of-speech tagging systems that was used in our work.

3.1. Historical overview of MT

Machine Translation (MT) refers to the "automatic translation of text from one human language into another" (Kenny, 2018:428). This implies that the entire translation process is conducted mostly without any human intervention, although they can intervene as pre- or post-editors or even to create data to train the systems. With the rapid development of technology and the increasing use of the internet globally, MT systems have been used increasingly in a variety of domains.

The first proposals in this field date back to the early 1930s, from Georges Artsrouni and Petr Trojanskij, who proposed the development of mechanical dictionaries that would allow searching for words in one language and, consequently, translating them into another. Trojanskij's proposal was more complex in the sense that he envisioned a multilingual translation mechanism consisting of three stages: firstly, a monolingual human translator would individually locate the words in a source text; secondly, the text would be translated by a machine, and finally, a second monolingual translator would correct the output (Hutchins, 1995). However, the systems and resources were limited, therefore, this proposal was not feasible for the time being.

However, the first efforts to automate translation emerged after World War II particularly due to the introduction of the first computers. As such, an important milestone in MT history occurred years after these proposals with the notable Warren Weaver Memorandum in 1949. Weaver was a former mathematician and, at the time of his memorandum, he was the director of the Natural Sciences Division at the Rockefeller Foundation. By having expertise in areas such as cryptography, mathematics and linguistics, he made several suggestions to overcome problems such as semantic ambiguity caused by the translation done at the word level, the translation done word by word. Among some of his proposals were the use of the context of occurrences of a word for semantic disambiguation. This would be done through statistical approaches that would allow deducing how much context would be necessary for the disambiguation (Kenny, 2018; Hutchins, 1995).

Weaver's Memorandum had a positive impact on the community, hence the following years were highlighted by the establishment of several research centers, not only in the United States (US), but in many European countries. Each group was researching different approaches, however, always drawing on linguistic research as a foundation for their systems. An important achievement occurred in 1954, with the first demonstration of a MT system carried out by Léon Dostert, a scholar from Georgetown University in collaboration with the International Business Machine Corporation (IBM). Although very limited, it successfully translated a few words from Russian to English.

However, even after some notable achievements done the years prior, during the 1960s some problems began to emerge that were difficult for MT to overcome, such as lexical ambiguity. The negative view of MT had a striking point in 1966, when a committee composed of US government investors was set to evaluate the progress done in the MT field. This was known as the Automatic Language Processing Advisory Committee (ALPAC). In their report they concluded that "MT was slower, less accurate and twice as expensive as human translation and that 'there is no immediate or predictable prospect of useful machine translation'" (Hutchins, 1995:435). Therefore, they did not see a need to keep investing in MT. Nonetheless, they recommended the development of tools that could support and help the work of human translators, such as automatic dictionaries, and tools that help them store important information. The ALPAC report had a negative impact on the research of MT in the US, therefore resulting in a slowdown in activity in the decade that followed it.

Despite this drawback, other countries had an increase of interest in this field. In Canada, English and French were established as official languages, which led to a growing demand for translation services for both languages. A noticeable system that was implemented was *Météo,* whose goal was the translation of weather forecasts from English to French (Kenny, 2018). In Europe, the Commission of the European Communities (CEC), the now European Union (EU), also saw an increased demand for translations of scientific, legal and technical documents for different languages, due to the inclusion of new member states (Hutchins, 1995). One of the first systems used was *Systran*, founded in 1968 by Peter Toma, already used in other organizations in the US.

During the 1980s and especially during the 1990s, a new turn in the MT field took place. Besides the already mentioned, many other commercial MT systems began to emerge, driven by the spread of personal computers and some other technological advances. Also word-processing tools were developed that would allow the post-edition of the text. In this period, many efforts were made to develop the MT systems and also extended them to other languages, such as Chinese and Arabic (Hutchins, 1995).

In the early stages of MT development, the focus was on achieving full automatic high quality translations (Hutchins, 2001). However, the 21st century has brought new demands. With the great technological advances and the increased use of the Internet, and consequently the more content available, one of MT's focuses is to be able to translate this content. Since the 2000s, more sophisticated systems that enable higher quality translations at a rapid pace are emerging. Two of these MT systems are Statistical MT and Neural MT, the latter being the current state of the art for MT. Both systems will be presented in detail in the next sections.

3.1.1. Rule-based systems

Rule-based machine translation (RBMT) systems were the first approach used in the first MT generations and were prevalent until the end of the 20th century. It consists of using linguistic knowledge, that is, it "relies on morphological, syntactic, semantic, and contextual knowledge about both the source and the target languages respectively and the connections between them to perform the translation task" (Shiwen and Xiaojing, 2014:186). Grammars and

dictionaries were used to map the translation of the words of a sentence and usually can be divided into three different types of systems: direct, transfer and interlingua.

The direct systems are characteristic of what was considered to be the first generation of MT. This approach consisted of translating a text by directly looking at the words using a dictionary. However, these systems were very limited, only integrating a few lexical rules, as well as rules that allowed local rearrangement of words, therefore, being insufficient to solve problems such as ambiguity or semantic problems (Kenny, 2018).

In order to overcome these limitations, the second generation of MT systems resorted to the transfer model. In this approach, the systems would analyze a source text and convert it into its syntactic internal structure. This internal representation is then converted into its equivalent representation in the target language via transfer. Lastly, with this representation of the target text, the systems generated a translated text (Shiwen and Xiaojing, 2014). However, because there were no limitations to how many rules there could be, these systems end up using a large set of rules that often contradict each other (Stein, 2018).

The last approach is the Interlingua. This approach consisted of using an artificial language, an interlingua, that was independent of any natural language and that would be able to represent information in any language. According to Kohen (2020:10), the goal of this system was to "(...) analyze a source sentence into its meaning, hopefully in a language-independent meaning representation called interlingua, and then to generate the target sentence from that interlingua representation." During the 1970s and 1980s several efforts were made to achieve an interlingua, however, not much progress was made and the interlingua is still far from concrete results still nowadays.

3.1.2. Data-driven systems

Data-driven systems, also referred to as corpus-based systems, emerged during the 1980s and became very popular by the turn of the century, thus replacing the RBMT systems. They were motivated especially due to the growth of the internet and consequently the increase of information available in different languages. Unlike the previous method, in this approach the systems learn from parallel corpora mostly. According to Kenny (2018:435), the "translation knowledge can be learned directly from parallel corpora (or 'bitexts'), that is, collections of

source texts aligned with their human translations". This implies that they translate without any explicit knowledge of linguistic rules. There are three types of data-driven systems: example based, statistical MT and neural MT.

The first data-driven models were the Example based systems (EBMT), emerging during the 1980s. The goal of this approach was "to find a sentence similar to the input sentence in a parallel corpus and make the appropriate changes to its stored translation" (Koehn, 2020:36). In a sense, these systems resemble a Translation Memory system.

Statistical Machine Translation (SMT) systems started to emerge during the late 1990s and were state-of-the-art until the mid-2010s. These systems learn from parallel corpora and in order to translate a text "it generates many thousands hypothetical translations for the input string and calculates which one is the most probable, given the particular source sentence, the models it has learned and the weights assigned to them." (Kenny, 2018:436) The most salient advantage of these systems is that they require no prior knowledge of a language in order to translate a text (Stein, 2018), but the lack of connection between n-grams, such as phrasal verbs, and the heavy dependency on frequency of the sentence had impact on the quality of the outputs.

Neural Machine Translation (NMT) is the current state of the art. These systems, like SMT models, are trained on high volumes of corpora, however they use artificial neural networks that resemble human neurons. These artificial neural networks are a "machine learning technique that takes a number of inputs and predicts outputs" (Koehn, 2017:6). These systems have an "encoder-decoder" architecture, where the encoder part of the system analyzes the source text and divides it into words, each of which are represented by vectors. The decoder part of the system, in turn, presents the most probable translation, for each representation position, taking into account the context in which it occurs (Forcada, 2017).

3.2. Translation Memory

In addition to the MT systems there are also other tools that mainly rely on the use of software to assist the translation process. These are called Computer-aided Translation (CAT), a set of different programmes aiming to facilitate the translator's work by enabling their work to be as effective as possible (Bowker and Fisher, 2010:60).

One of the main tools within a CAT environment is the Translation Memory (TM). As aforementioned, these systems have the property of storing previously translated sentences or segments. They were initially created to assist translators, however, they are often used in conjunction with MT. They differ from MT in the sense that TM systems only allow to store and retrieve previous translation segments and the translator will decide if the proposed TM should be used or not (Garcia, 2009). Also, if there is not a match for a specific segment, the MT will be used to translate this new segment.

3.2.1. TM history

Despite its negative impact, the already mentioned ALPAC report was an important booster for the development of CAT tools. In the report, the committee points out the positive example of two institutions in Europe, where machines were being used to assist the translators. The first one was the Federal Armed Forces Translation Agency, in Germany, who used computer systems for the retrieval of glossaries. Also, in the European Coal and Steel Community (ECSC), the now European Union, there were "(...) developed and used a computer system to retrieve terms and their contexts from stored human translations by identifying those sentences whose lexical items most closely matched the lexical items of a sentence to be translated" (Reinke, 2018:56). However, in both places, the translations were done by humans and not by a MT system. Even so, the committee encouraged the development of computer based systems that could be of assistance to human translators (ALPAC, 1966).

One of the first formal proposals of a TM system was made by Peter Arthern, a representative of the Council of the European Communities. In a paper published in 1979, he discussed the role of human translators and the computer and also "the potential use of computer-based terminology systems in the European Commission." (Hutchins, 1998:294). He also proposed a concept that he named "translation by text-retrieval", which he described as follows:

> *The pre-requisite for implementing my proposal is that the text-processing system should have a large enough central memory store. If this is available, the proposal is simply that the organization in question should store all the texts it*

> *produces in the system's memory, together with their translations into however many languages are required.*
> *This information would have to be stored in such a way that any given portion of text in any of the languages involved can be located immediately, simply from the configuration of the words, without any intermediate coding, together with its translation into any or all of the other languages which the organization employs.* (Arthern, 1979:93,94)

In his paper, Arthern mentioned that many of the translated documents were often repeated and some would even be quoted from previous translations. Therefore, he proposed the creation of a text processing system with a large storage capacity that would store texts and their translations in all languages that were necessary. This would allow for easy navigation between texts, thus enabling human translators to save time while translating.

Another significant proposal was made by Martin Kay. In his paper "The Proper Place of Men and Machines in Language Translation", published in 1980, he proposed what he named "The Translator's Amanuensis". This proposed system would be a text processing system that would have two functionalities. On one hand, it would function as a dictionary that would allow the translator to press on any word in the text and automatically retrieve the meaning of the word. On the other hand, this system would allow the translator to also analyze the text and search for previous translations that would be similar to the current one. As described by Kay (1997):

> *If the piece of text to be translated next is anything but entirely straightforward, the translator might start by issuing a command causing the system to display anything in the store that might be relevant to it. This will bring to his attention decisions he made before the actual translation started, statistically significant words and phrases, and a record of anything that had attracted attention when it occurred before. Before going on, he can examine past and future fragments of text that contain similar material.* (Kay, 1997:19)

According to Kay's proposal, a translator would have both the text to be translated and its translation, however, they would be able to select and edit the relevant text. His main goals was the development of a system that would increase the translator's productivity and, overall, to decrease their dependence on MT.

Both proposals had an important impact in the development of Translation Memories. They emphasized the importance of having tools to assist translators, allowing them to be more productive in their work.

These systems got their name exactly because they act as memories or archives (Reinke, 2018). They became commercially available during the 1990s. They became part of the *translations workstation* or *translation workbench*, a set of other machine-aided human translation tools that translators could use. Some examples are the Automated Language Processing System (Alps), Trados and IBM's corporation TranslationManager/2.

3.2.2. Types of TMs and limitations

As aforementioned, the TM retrieval is done by matching a source text to other stored segments. Two types of matching were already mentioned in the previous chapter (see section 2.2.1), namely exact and fuzzy matches, however, there are several other different types of segment matching in TMs. Bowker and Fisher (2010) also distinguish full match, term match, sub-segment match and, lastly, no match.

A *full match* occurs whenever "a new source segment differs from a stored TM unit only in terms of so-called variable elements, which are sometimes referred to as "placeables" or "named entities"" (Bowker, 2002:98). These variable elements can be dates, times or currencies. The *term match* is commonly used for matching specific terms. In this case, the terms in a source text have to match the ones in the system. The *sub-segment match* is an intermediate match between the last two matches. In this case, only a sequence of adjoining segments are identical to the original text. Lastly, a *no match occurs when* no compatible translations to the new text are found. In this case, the new text can be added to the database (Bowker and Fisher, 2010).

Although these systems are very useful for the translation process, allowing among others, the reduction of not only the time spent on the translation but also of the costs associated

with it, and overall maintaining the consistency of translations, TM systems have some limitations.

One of the limitations that can be pointed out is that these systems depend not only on the quantity, but also on the quality of the stored segments. Bowker (2002) states that low quality segments can compromise the quality of the text as a whole. A second limitation is related to the fuzzy match. The boundary between this and other types of matches is not always straightforward. Also, the process of post-editing a TM is not simple, usually, they are reviewed manually by a human editor or even by several editors, in order to ensure their quality. Another limitation is the fact that the systems do not always allow saving more than one option for the same TM entry.

Another drawback of these systems, and particularly important for the current work, is the fact that these systems operate at sentence level. This can have two implications. On one hand, this can lead the translator to discard inter and intrasentential dependencies, only focusing on the current sentence, while editing the segment (Reinke, 2018). On the other hand, the segments being isolated from context may result in contextual problems, such as ambiguity, register and gender issues.

3.3. MT, TMs and context

The Neural Machine Translation systems have revolutionized the way translation is handled. Until the introduction of these systems, most of the MT systems were based on *n-gram* models. These models "reduce the probability of a sentence to the product of word probabilities in the context of a few previous words" (Koehn, 2017:32). Thus, this implies that the basic unit for translation was the sentence. In this type of approach, when only the translated sentence, isolated from context, is considered, it may not display grammatical or coherence problems. However, when considering the entire document, the same results usually are not obtained (Hassan et al., 2018). Only recently, the systems are capable of integrating more information beyond a sentence, tackling even a full document. This new approach is called *document-level translation* or *context-aware machine translation* and integrates more information beyond the sentence boundaries by incorporating linguistic discursive phenomena (Lopes et al., 2020; Yin et al., 2021).

Both TM and MT systems are complementary components. Whenever there is a match between a source text and segments stored in the database, those are used, therefore, in such cases there is no need for MT systems. However, if this does not happen, the segment is translated by the MT system, which is often specifically trained for each client. In this sense, TMs are important to enhance the effectiveness of the translation process and are used if there is already a match for it. However, TM systems also operate at a sentence level. This can have negative effects on the translation quality. Selecting segments, even with a high level of compatibility, can create problems of cohesion and coherence in the document and other issues, such as lexical ambiguity, agreement and referential problems, are very common. This concept is very troublesome for linguistics. A text is seen as a unit of meaning and the sentences in a text establish a relationship of dependence and meaning.

Context-aware MT aims to improve common problems in translation. One of these is the ambiguity that is inherent to all natural languages. Ambiguity manifests itself lexically, since words often have more than one meaning, hence, can have more than one possible translation and the most adequate word choice often is context-dependent. It can also be due to structural ambiguity since there can be multiple meanings for a sentence (Koehn, 2020).

3.3.1. Definition of context

In order to achieve a good translation, it is necessary to preserve the meaning of a source text. Therefore, the translated text needs to be fluent and completely correct (Koehn, 2020). However, this process is not always simple. We stated previously that the major problem that is faced in translation is the ambiguity of languages. Words and even the structure of a sentence can be ambiguous and can have more than one possible translation. Therefore, to understand the meaning of a sentence or a text it is necessary to understand their context. Context as a concept has long been the subject of various disciplinary areas and, consequently, its definitions can vary substantially. To define it, some authors use various dimensions (House, 2006). Two specific dimensions of context are particularly important to both translation studies and linguistics: *linguistic contex*t and *extralinguistic context*.

Linguistic context, also designated as co-text, can be described “by surrounding text within a particular version of one document but not limited to the current sentence” (Melby and

Foster, 2010:6). It refers to all linguistic elements present in a text that allows one to understand the full meaning not only of a sentence, but also of the text itself (Yule and Widdowson, 1996).

In contrast, extralinguistic context or situational context refers to the extralinguistic aspects that complement a text. The term was coined by Malinowski, an anthropologist, in a publication from 1923. After observing the native speakers from the Pacific and Trobriand Islands, he noticed that the same words or expressions can have different meanings according to the context in which they were used. Consequently, he pointed out that the meaning of a word cannot only be defined according to linguistic information but also the communicative situation (Malinowski, 1923). According to Halliday (1985), this concept comprehends three aspects: *field*, *tenor* and *mode*. The first aspect, *field*, refers to what is happening between the participants of a communicative event. *Tenor* refers to the participants themselves, that is, the roles they play in the conversation and the kind of relationship that they have with each other (closeness or distance). Lastly, *mode* can be described as "what part the language is playing, what it is that the participants are expecting the language to do for them in that situation (...)" (Halliday and Hasan, 1985:12). This phenomenon, then, involves all the non-linguistic information that contributes to the correct interpretation of the text.

Taking these two concepts into consideration, for the present thesis, we consider that context refers to all the information that proceeds and follows a given segment or sentence and is mandatory to understand the meaning beyond a single sentence, which involves all the linguistic elements, as well as the extralinguistic knowledge. Thus, context provides the necessary information to the correct interpretation of a phrase that otherwise would not be correctly understood and to disambiguate any ambiguities.

In translation, the problems related to the lack of context are associated with discourse phenomena. For this purpose a distinction can be made between text and discourse. While a text can be defined as "the verbal record of a communicative event" (Brown and Yule, 1983:190), this is, a linguistic unit that stems from a communicative intervention, the term discourse refers to the language usage in context, meaning, in the way that it is used by the speakers. The latter can be described as "the product of the use of grammar in particular natural contexts" (Ariel, 2009:5), and it presents a sequential and cohesive organization of sentences. Discourse phenomena are thus all the linguistic elements present throughout the text that allows the

organization of ideias and the reference of entities. This phenomena can be divided into two categories: *cohesion* and *coherence phenomena*.

3.3.1.1. Cohesion phenomena

Cohesion refers to the semantic and syntactic relations that exist within a text and that are important for its correct interpretation. The relationship between elements on a text can be referred to as cohesive chains, this is "a semantic relation between an element in the text and some other element that is crucial to the interpretation of it" (Halliday and Hasan, 1976:8). These cohesive ties are not strict to the sentence level, hence, can be established and maintained at different spans. This type of phenomena is related to the linguistic context and manifests itself through pronominal anaphora and lexical cohesion.

Anaphora is a reference phenomena that "refers back to a concept or word presented in a prior sentence, the concept referred to is "old" information and serves as a marker for knowledge already possessed by the reader" (Fishman, 1978:160). Anaphoric references do not occur only within the sentence, rather this reference can occur between sentences. For this reason, this linguistic phenomena has been the subject of substantial research in MT (Mitkov et al., 1995; Voita et al., 2018). The research has focused on pronominal anaphora. It is the use of personal pronouns, usually from the third person to refer to entities previously mentioned, which are called *antecedent*. In MT, anaphora resolution is a difficult task mainly due to ambiguity, since it is not always possible to define the antecedent in a referential chain.

Lexical cohesion "is the cohesive effect achieved by the selection of vocabulary" (Halliday and Hasan, 1976:274). It is often achieved by two mechanisms: repetition and collocations. Repetition refers to the reiteration of a lexical item using processes such as synonyms or by a superordinate (Halliday and Hasan, 1976), thus avoiding excessive repetition of a same word or expression. As for collocations, it refers to the use of "related words that generally co-occur" (Maruf et al., 2019:5). This phenomenon ensures that both the repeated elements and collocations are consistent throughout a text.

3.3.1.2. Coherence phenomena

Different from the previous, coherence refers “to the relationship between sentences that makes real discourses different than just random assemblages of sentences” (Jurafsky and Martin, 2003:442). This is an underlying mechanism that ensures the logical sequence of ideias, not only within a sentence but also between all elements from a text. This phenomenon involves much more than grammatical elements. It also involves the speakers’ knowledge of the world. In machine translation, it is related to the discourse connectives.

Discourse connectives or discourse markers are often functional words “that signal the existence of a specific discourse relation or discourse structure in the text.” (Maruf et al., 2019:6) These connectors establish relations such as cause or contrast between segments. This is the case of the words such as *however* or *yet*, which allows to express contrast between clauses or the word *moreover*, that allows the addition of more information. According to Meyer and Webber (2013), what constitutes a problem is the crosslinguistic variation of the source language into the target language, which can also be ambiguous.

The notion of context is important for translation as it allows the interpretation of a translated text, making it possible to handle problems that manifest at a document level. It is also important to eliminate ambiguity.

3.3.2. Context-related typologies

As mentioned previously, context-aware MT integrates information beyond the sentence boundaries by incorporating discursive phenomena. To investigate this phenomenon it is common to use test suites with the subjects in question to test the performance of the system.

Guillou et al. (2018) analyzed and evaluated the performance of 16 NMT systems on the translation of pronouns from English to German with a test set with 200 pronouns.The authors found that all of the NMT systems analyzed had a better performance translating pronouns with intersentential reference. In contrast, the translation of anaphoric pronouns, whose reference is intrasentential, were more difficult.

Bawden et al. (2018) created English to French test sets that tackled coreference, lexical coherence and cohesion as context-aware categories, in order to test the performance of a NMT system with a multi-encoder architecture. The results for their experiment showed positive

outcomes for both coherence and cohesion, and conversely, less favorable results for coreference.

Voita et al. (2019) analyzed the performance of NMT systems regarding context-aware phenomena by creating two different datasets. The first one consisted of English to Russian subtitles to identify three types of inconsistencies in the source text: deixis, ellipsis and lexical cohesion from English to Russian translation, in order to improve evaluation metrics for document-level translation. Their second dataset consisted of a larger sentence-level corpora that was contrasted with another one aligned at a document level.

Cai and Xiong (2020) examined pronouns, discourse connectives and ellipsis in English to Chinese test suits in different NMT systems in order to test their performance. Overall, the results obtained were positive for both discourse connectives and pronouns, however, ellipsis was the problematic category for the systems.

Yin et al. (2021) analyzed pronouns, ellipsis, lexical consistency, formality and verb forms as context-aware issues in a parallel corpora of 14 language pairs in order to evaluate the performance of an MT model. The authors concluded that ellipsis was the discourse phenomena the most problematic categories, however, they noted that the context-aware MT models improved.

In Castilho et al. (2021), the author annotated an English-Brazilian Portuguese corpus made of 60 documents with a total of 3,680 sentences, from six different domains: literary, subtitles, news, reviews, medical and legislation. They consider gender agreement, number agreement, lexical ambiguity, reference, ellipsis and terminology as context-aware issues.

3.4. Part-of-Speech

Parts-of-Speech (POS), typically referred to as word class, enables one to understand how words function within a sentence (Carnie, 2012). Commonly, there are eight categories distinguished: verb, noun, pronoun, adjective, adverb, preposition, conjunction and interjection. All of the aforementioned categories are divided into two different classes: open and closed classes. An open class, as suggested by the name, allows new words to be added to it. Categories such as nouns, verbs and adjectives belong to this class type. In contrast, closed classes have a defined number of members and the coinage of new forms are rare or nearly nonexistent. This is

the case of prepositions and conjunctions which are also often categorized as functional words, since their role is purely grammatical having no semantic meaning.

However, grammatical categories are not only determined according to their meaning. They are defined according to their morphological and syntactic distribution. Morphological distribution refers to the affixes (i.e. inflectional or derivational) that can be attached to certain words. In the latter, syntactic distribution refers to words that typically can co-occur with each other (Carnie, 2012).

Part-of-speech tagging or POS tagging is a sequence labeling task, a common sub-task in NLP, which corresponds to the attribution of a tag or a label to a word sequence (Jurafsky and Martin, 2020). These systems have a sentence as a basic unit. The sentence is tokenized, this is, divided into words, and to which word it is attributed a tag. A tagset corresponds to a list of grammatical categories previously established and adapted to the system (Güngör, 2010).

There are several POS tagging systems, however for this project we used one already used in the company, namely Stanza[3]. This system is a set of tools that enables natural language analysis and processing developed by the Stanford NLP Group and supports more than 70 languages. It has a neural pipeline and has many tools such as a tokenizer and a lemmatizer, however, for our work, only the POS tagger was used. Stanza uses a text as an input and produces POS tags as outputs, such as verb, noun, pronoun, or adjectives (Qi Peng, et al., 2020). The tagset that is used in this system is Universal Dependencies[4] (UD). This is a multilingua framework that has a total of 17 tags, along with all relevant morphological features such as gender, case, animacy and many others. *Figure 4* shows the dependency tags used in the POS system:

[3] "Overview - Stanza - Stanford NLP Group." https://stanfordnlp.github.io/stanza/. Accessed 23 Mar. 2022.
[4] https://universaldependencies.org/. Accessed: 17 Aug. 2022.

Traditional POS	UPOS	Category
noun	NOUN	common noun
	PROPN	proper noun
verb	VERB	main verb
	AUX	auxiliary verb or other tense, aspect, or mood particle
adjective	ADJ	adjective
	DET	determiner (including article)
	NUM	numeral (cardinal)
adverb	ADV	adverb
pronoun	PRON	pronoun
preposition	ADP	adposition (preposition/postposition)
conjunction	CCONJ	coordinating conjunction
	SCONJ	subordinating conjunction
interjection	INTJ	interjection
–	PART	particle (special single word markers in some languages)
–	X	other (e.g., words in foreign language expressions)
–	SYM	non-punctuation symbol (e.g., a hash (#) or emoji)
–	PUNCT	punctuation

Figure 4 *- Universal part-of-speech tags (UPOS) extracted from Marneffe et al. (2021)*

4. Methodology

This chapter will present the description of the experiments conducted, the data collected and analyzed and the methodological procedures applied throughout the course of the internship.

As mentioned previously, the goal of the internship was to create gender neutral and context-independent segments by using POS information to automatically identify context-dependent segments. Therefore, the internship focused on two major tasks. The first one consisted of analyzing two datasets of TMs in different languages: one with segments that had been previously marked as context-dependent by the community of terminologists and senior editors; and another dataset with the context annotation done by us. The goal of this analysis was to understand what linguistic clues promoted context-related issues and if those same clues would be similar across languages.

The second task consisted on using a POS tagger in all segments that were previously annotated by us, both context-dependent and context-independent. For this, we selected four language pairs and ran a POS tagger in the data to check for morpho-syntactic patterns. The goal was to use the POS information obtained in the previous task and co-relate specific patterns with context-dependent TMs vs. context-independent ones.

4.1. Pilot experiment - Context-dependent TMs

The first experiment consisted of analyzing a sample of curated TMs, extracted from the TM server, that had been previously marked as context-dependent by terminologists or senior editors in the curation process.

As aforesaid, Unbabel works with a community of terminologists and senior editors who work directly with the machine translated content. They are often involved in different tasks, including the curation of TMs. During this process, terminologists are asked to correct errors in the segment, if any, and also to identify whether or not the analyzed segment is context-dependent. To assist them during the task, they are given guidelines with examples of common context-related issues. When classifying, terminologists/senior editors only have access to the TM as an isolated segment and not as a part of an entire message. This means that they are performing this task without any access to the context.

Classifying context was not a regular task of the curation process and represented a first experiment to test the impact of having context dependent TMs, segments that would not be blocked from edition, allowing the community of editors to perform the necessary changes according to the context, vs. context-independent TMs, segments that could be blocked from edition, allowing the editors to spend less time checking for non-existent errors.

Therefore, the goal of this analysis was to validate this classification done by terminologists/senior editors, in this preliminary pilot, to understand if the classified segments were in fact context-dependent and if the design of this experimental task needed further improvements. For this analysis, the language pairs selected were Brazilian Portuguese (PT-BR) and Spanish (ES), English being the source text for both. The dataset comprised 360 TMs for PT-BR and 13,295 TMs for ES (see *Table 1*).

As expected, many of the TMs contained in the original dataset were repeated, that is, more than one entry for a single TM, thus all the duplicates that occurred more than once in the dataset were removed to ensure that only unique entries would be considered. As a result, the final dataset comprised 45 unique TMs entries for PT-BR and 9,296 unique entries for ES. For ES, however, due to time constraints, it was not possible to analyze all the segments, hence a random portion of the final dataset was chosen for analysis, corresponding to 2,000 out of the 9,296 unique TMs.

	Unique entries	Removed TMs	Total TMs Extracted
PT-BR	45	315	360
ES	9,296 (2,000)	3,999	13,295
Total:	2,045	4,314	13,655

***Table 1**. Pilot experiment dataset*

Altogether, 2,045 segments were analyzed for the first experiment (see *Table 1*). At the time of this experiment, the task of annotating context-related issues was newly introduced in the company, which is reflected by the disparity of data available for both language-pairs.

4.3. Context-dependent TMs annotation guidelines

After the pilot experiment, namely the validation of the context-dependent TMs annotation done by terminologists/senior editors, it was necessary to improve the annotation guidelines and develop a typology that would cover all of the contextual problems that were encountered. As stated in the previous chapter (see section 3.3.2), in the literature, there are several proposals of typologies to account for common context-aware problems that often coincide in a core set of categories, anaphoric pronouns and ellipsis being the common ones amongst all (Guillou et al., 2018; Bawden et al., 2018; Voita et al., 2019; Cai and Xiong, 2020; Castilho et al., 2021, and Yin, et al., 2021).

Although there are several proposals, we based ourselves on the typology of Castilho et al.(2021). The authors present a significant number of categories that were suitable for the context related issues that were found in the first experiment. Therefore, our proposed typology for context-related issues comprises the following categories: gender agreement, number agreement, register, ellipsis and terminology.

4.3.1. Gender agreement

English is a language with no grammatical gender. However, many languages have it. This implies that words in these languages can have either feminine, masculine and neutral forms. There are specific word classes such as adjectives (also participial forms of the verbs), articles, names and pronouns, that require gender markers. Thus, gender agreement implies the correlation between the gender of the pronoun, name or adjective and the word that it relates to, which can be in a different sentence. The following examples illustrate gender constraints.

- Context-dependent:

(1) EN: Thank you for contacting us.
PT**: Obrigado** por entrar em contacto connosco.

- Context-independent:

(2) EN: Thank you for contacting us.
PT**: Agradeço** por entrar em contacto connosco.

In the example (1), the word in bold has a masculine gender marker in the participial form of the verb. This form is valid if the receiver/addressee of the message is a male, but it would not work if the receiver is a woman. Therefore, the segment is context-dependent. However, in the example (2), the sentence is gender neutral, that is, the verb (with the same meaning, but not a participial form) chosen does not need gender markers, therefore the segment is independent of context.

For our analysis, gender agreement was noted whenever it was not possible to disclose the gender of the referent through the segment.

4.3.2. Number agreement

Similarly to the previous category, number agreement implies the correlation between the number of a pronoun, name or adjective and the word that it relates to, which can be in a different sentence. There is a number disagreement issue whenever there is a disagreement between linked/indexed elements in a sentence. This can be seen in the following examples:

- Context-dependent:

(1) John really liked the **product**. (...) **They** were very high quality.

- Context-independent:

(2) John really liked the **product**. (...) **It** was very high quality.

4.3.3. Ellipsis

In the broad sense, Ellipsis is a linguistic phenomena that refers to the omission of one or more words or expressions from a sentence. The omission of information from a clause may compromise the comprehension of it. In this case, a broader context is needed to understand the sentence. Usually the information that proceeds the clause helps disambiguate the meaning, therefore it implies that ellipsis is dependent on the previous clause (Castilho et al. 2021).

- Context-dependent (informal register):

(1) EN: Any time that you make a change in your account, **even if it's a photo,** we will send you an email.

PT: Sempre que efetuas uma alteração na tua conta **[∅]** nós iremos enviar-te um e-mail.

In the example, part of the text is missing from the segment and it affects the interpretation of the original message.

- Context-independent (informal register):

(2) EN: Any time that you make a change in your account, **even if it's a photo**, we will send you an email.

PT: Sempre que efetuas uma alteração na tua conta, **mesmo que seja uma foto**, nós iremos enviar-te um e-mail.

In these examples, the omitted parts of the previous examples are now restored and the sentence is now correct and intelligible.

4.3.4. Terminology

Terminology refers to a vocabulary of a specific domain (e.g. tourisms, tech, retail, etc.). This category is very important as the company works with Glossaries. This tool is a compilation of specific words and phrases given by the company's clients. A specific-domain term or glossary word is context-dependent when is wrong selected and not the right choice for that context. It is important to make sure that it is correctly translated.

- Context dependent:

(1) EN: Thank you for contacting our customer support.

ES: Gracias por ponerse en contacto con nuestro servicio de **Soporte al Cliente**.

In the example above, *Customer Support* is a glossary term, therefore the expression is a specific-domain term. Although it is a viable alternative, it is not compliant with the request form the client.

- Context independent:

(2) EN: Thank you for contacting our customer support.

ES: Gracias por ponerse en contacto con nuestro servicio de **Atención al Cliente**.

Unlike the previous example, the term contained in the glossary is well applied and therefore, it is not mistaken for other alternatives not compliant with the client's requirements.

4.4. Context annotation in TMs

The current experiment consisted of analyzing TMs not as isolated segments but integrated into a full text, namely an email thread, using the newly proposed context-dependent TMs annotation guidelines (see previous section). Tickets that are translated by the company have a significant proportion of TMs. For the selection of the corpus to be annotated, a batch of customer support tickets composed almost exclusively of TMs was selected for analysis, in order to have representative data. This was important because most customer support emails have a substantial portion of templated content, hence most of the tickets already contain multiple TMs.

As for language pairs, the analysis was done for European Portuguese (PT), Brazilian Portuguese (PT-BR), European Spanish (ES) and Latin-american Spanish (ES-LATAM), English being the source text for all the LPs. The dataset comprised 6,368 TMs for PT-BR; 28,604 TMs for ES; 33,623 TMs for PT-BR and 10,026 for ES-LATAM. Due to the high volume of entries, only a sample of the original dataset was selected for analysis: 1,300 TM entries for all language-pairs (see *Table 2*). After selecting a portion of the original dataset, the following step was to analyze all the data and identify the TMs that generated context-related issues at a document level.

	Total TMs entries	Total TMs analyzed

PT	6,369	1,300
PT-BR	33,623	1,300
ES	28,604	1,300
ES-LATAM	10,026	1,300
Total:	78,622	5,200

***Table 2**. Second experience dataset*

After all the TMs with context-related issues were identified, the following step was to extract POS categories by using a POS tagger system. Our hypothesis was that by using a POS tagger we could extract grammatical patterns to identify context-dependent segments, specifically the ones with gender.

4.4.1. POS tagging

For this analysis, we used Stanza, the POS tagger already used in-house (see section 3.4). The goal of using Stanza was, firstly, to identify what grammatical categories were involved in context-dependent segments and, secondly, to understand if there were patterns (sequences of POS) that could distinguish context-dependent TMs from context-independent ones. Therefore, for this experiment, we analyzed all the sentences that were annotated as context-dependent, 335 segments, and analyzed the exact same number of context-independent segments. This allowed us to have a balanced dataset and to examine the POS patterns of both context-dependent and context-independent segments. *Table 3* shows the exact number of segments selected for all LPs.

	Context-dependent TMs	Context-independent TMs	Total:
EN-PT	102	102	204
EN-PT-BR	75	75	150
EN-ES	78	78	156

EN-ES-LATAM	80	80	160
Total:	335	335	670

Table 3. *Sampled dataset for POS analysis*

4.5. POS patterns in context-dependent vs. context-independent TMs

One of the goals of this research was to survey sequences of POS patterns that would identify context-dependent segments with gender constraints. Therefore, for this final experiment, we wanted to validate the POS patterns found in the context-dependent and context-independent TMs analyzed (see section 4.4.1). For that, we selected a new dataset, also consisting of isolated segments, that is, the TMs were not in context, from seven different clients with high volumes of TMs from different domains. We then ran Stanza in this data, and did a manual evaluation that matched the POS patterns with the context (dependent vs. independent). In total, this dataset comprised 8,000 TMs, 2,000 for each LP. As for language pairs, the analysis was done again for PT, PT-BR, ES and ES-LATAM, English being the source text for all the LPs (see *Table 4*).

	Clients:	**Domain:**	**Total of per client:**
PT	Client 1	Technology	1,555
	Client 2	Gaming	445
PT-BR	Client 3	Gaming	2,000
ES	Client 4	Retail	923
	Client 5	Gaming	100
	Client 6	Retail	977
ES-LATAM	Client 7	Gaming	2,000
Total:			8,000

Table 4. *Experiment 3 dataset*

5. Results

This chapter will present the results obtained from the following experiments conducted during the course of the internship: for the pilot experiment (section 5.1), we analyzed TMs that had already been classified by the community of senior editors and terminologists as context-dependent, in order to understand the types of context-related issues found; for the second experiment (section 5.2), we performed that classification ourselves in a new dataset, using the designed context-dependent TMs annotation guidelines. We also performed a POS tagging step in this data, in order to find POS patterns that would distinguish context-dependent and context-independent TMs. Finally, we validated the POS patterns found in a new and unclassified dataset from seven clients from different customer support domains.

5.1. Pilot experiment - Context-dependent TMs

For the pilot experiment, we analyzed a total of 2,045 segments for PT-BR and ES. These were previously annotated as context-dependent by senior editors and terminologists. Results were divided into two categories: “true positives” and “false positives”. The first category corresponds to segments that in fact needed context. The latter category corresponds to the data that was erroneously classified, namely TMs that were classified as being context-dependent even though they were not. The detailed results can be found in *Table 5*:

	Total of TMs	True Positives	False Positives
PT-BR	45	37 (82%)	8 (18%)
ES	2,000	106 (5%)	1,894 (95%)
Total:	2,045	143 (7%)	1,902 (93%)

***Table 5**. TMs analyzed*

These results show some disparity in the classification done by the terminologists/senior editors. On one hand, for PT-BR, only 18% of the data was misclassified. However, for ES, a very substantial portion of the data was misclassified (95%). Further on we will present possible explanations for such results obtained.

As for the context-related issues found for both language pairs there were mainly two: gender agreement and register. *Table 6* shows the exact number of context issues found per language pairs.

	Gender agreement	Register	Total:
PT-BR	37 (100%)	0	37 (26%)
ES	78 (74%)	28 (26%)	106 (74%)
Total:	115 (80%)	28 (20%)	143

Table 6. *Context-related issues for the pilot experiment*

Gender agreement accounted for the majority of the context related issues representing 100% of the cases for PT-BR and 74% for ES. Gender agreement implies the correlation between a masculine or a feminine word and information from previous or following segments. Examples (1) and (2) show instances of segments with gender constraints.

(1) EN: I'm really sorry to read you were dissatisfied with our previous reply.
PT-BR: Lamento muito saber que você ficou [**insatisfeito**] com a nossa última resposta.

(2) EN: I'll be more than happy to assist you and find out the information for you.
ES: Estaré [**encantado**] de ayudarte y averiguar la información para ti.

Both Portuguese and Spanish are languages that have grammatical gender, meaning that words can either have a feminine or a masculine form. What poses a problem in the examples above is the adjectives *insatisfeito* in (1) and *encantado* in (2) that are both in their masculine form. Even though the sentences are grammatically correct, due to the presence of the adjectives, these segments would not work if the addressee was a woman.

As for Register, it corresponds to the tone used in the text, namely, formal or informal register. Therefore, register errors involve the misuse of forms of treatment.

(3) EN: Hope you are well. [informal register]

ES: Espero que [**esté**] bien.

(4) EN: Hi! [formal register]
ES: Hola:

Both segments in (3) and (4) were misclassified regarding their register. This will result in coherence problems at the document-level as they do not respect the register required by the customer.

Overall, there was a discrepancy between the results obtained. As aforementioned, this task was newly introduced at the time of the analysis and represented a first experiment in context classification in the company. Also, the fact that terminologists and senior editors were not familiar yet with the task may have resulted in different classifications for it. It is important to note that the guidelines provided were very general and left some room for ambiguity. On one hand, the Brazilian Portuguese terminologists and senior editors seem to be more conservative, in the sense that less errors were found in the classification of context. On the other hand, the results obtained for Spanish showed a tendency to over-annotate.

One of the following steps after the analysis was to reformulate the previous guidelines, making them more concise and less ambiguous, in order to provide clearer instructions of what was intended with the experiment. As presented in section 4.3, we described in detail the specific categories that should be considered as context-dependent, as well as objective examples and counterexamples for each one of them. We also tested the proposed guidelines when annotating TMs as context-dependent or context-independent in the second experiment, context annotation in TMs, described in section 4.4.

5.2. Context annotation in TMs

The results of the pilot experiment not only showed that the guidelines needed improvement, but also that a new approach should be used. Therefore, the current experiment did not involve the community of terminologists' and senior editors annotation in segment TMs, but rather an annotation performed by us of context related issues in tickets with full context, in order to have context information. The dataset comprised a total 5,200 TMs, 1,300 segments for

LP. Results showed that 93.5% of the TMs were context-independent and 6.5% were context-dependent (see *Table 7*).

	Context-independent TMs	Context-dependent TMs	Total:
PT	1,197 (92%)	103 (8%)	1,300 (25%)
PT-BR	1,224 (94%)	76 (6%)	1,300 (25%)
ES	1,222 (94%)	78 (6%)	1,300 (25%)
ES-LATAM	1,219 (94%)	81 (6%)	1,300 (25%)
Total:	4,862 (93.5%)	338 (6.5%)	5,200

Table 7. *Second experiment results*

It is important to mention that, for this analysis, we used the typology proposed in section 4.4. The context-related issues found in this dataset were gender agreement, register, ellipsis and terminology (see *Table 8*).

	Gender agreement	Register	Ellipsis	Terminology	Total:
PT	101 (98%)	1 (1%)	-	1 (1%)	103 (31%)
PT-BR	75 (99%)	-	1 (1%)	-	76 (22%)
ES	77 (99%)	1 (1%)	-	-	78 (23%)
ES-LATAM	78 (96%)	2 (2%)	1 (1%)	-	81 (24%)
Total:	331 (97.9%)	4 (1.2%)	2 (0.5%)	1 (0.3%)	338

Table 8. *Second experiment context-related issues per category*

Similar to the pilot experiment, the vast majority of the context-dependent segments found in this analysis were marked by **gender** related issues for all language pairs, accounting for 98% of the total cases. As aforementioned, gender agreement implies the correlation between a masculine or feminine forms (i.e. pronouns, names or adjectives) and information from

previous or following segments. This category seems to be the most problematic category for every language pair and also highlighted in the two experiments.

(5) EN: We will try to help you.
PT: Vamos tentar ajudá-[**lo**].

(6) EN: Until then, take care, stay safe and have a great day!
ES: ¡Hasta entonces, cuídese, manténgase [**seguro**] y que tenga un buen día!

The issue with these segments relies on the pronoun *-lo* in (5) and the adjective *descansado* in (6). Both of them are in the masculine forms and, therefore, cannot be used in every context, which implies that they are dependent on contextual information located beyond the sentence level, hence these TMs do not work in every context because of the gender.

In this dataset, it was noticeable the clear tendency to use masculine forms as opposed to feminine forms (see *Table 9*). When analyzing the distribution of both forms, it is noticeable that the masculine is predominant marking 98% of the data. This may be an indication that masculine was being used as the “default" or as the “generic” gender.

	Masculine forms	Feminine forms	Total:
PT	98 (97%)	3 (3%)	101 (31%)
PT-BR	74 (99%)	1 (1%)	75 (23%)
ES	76 (99%)	1 (1%)	77 (23%)
ES-LATAM	78 (100%)	-	78 (23%)
Total:	326 (98%)	5 (2%)	331

Table 9. *Distribution per gender in experiment 2*

As opposed to the previous analysis, where **Register** marked 20% of the cases, in the current experiment, it only comprised 1.2% of the context-dependent segments. PT-BR was the only language that did not have an instance of this category. Again, this category is related to the

use of pronouns or forms of treatment that are not being used in the register required by the client (examples 7 and 8).

(7) EN: Thank you for contacting URL-0. [Informal register]
PT: Obrigado por [**entrar**] em contacto com a URL-0.

(8) EN: We hope that you continue to enjoy playing our game. [Formal register]
ES: Esperamos que [**sigas**] disfrutando de nuestro juego.

The segment in (7) is formal and the one in (8) is informal. However, they were identified with the incorrect register, therefore they needed to be considered context-dependent since they will cause coherence problems in a document.

The last category is **Ellipsis**, which accounted for 0.5% of the cases, however it only occurred in PT-BR and ES-LATAM. It refers to the omission of information (i.e. words, expressions) within the segment and affects the correct comprehension of the segment or may create ambiguities, as seen in the following examples:

(9) Previous segment: However, in order to proceed further, we will require a set of details from you.
EN: They are as follows:-
PT-BR: Eles são os seguintes: - [**Ø**]

Following segment: Looking forward to your reply.

(10) Previous segment: For further proceedings, we would request you to please elaborate on your issue so that we can assist you accordingly and provide you with a better solution.
EN: Once, we receive the information.
ES-LATAM: Una vez recibamos la información, [**Ø**]

Following segment: We will be glad to assist you further.

In the examples (9) and (10), the problem relies on the omission of the directives given by the agent. This is certainly a result from an error on the agent side, however, it compromises the meaning of the full text.

As in the previous analysis, gender agreement remained the prevalent context-related issue in this dataset with 98% of the total results. In addition, two new categories that had not occurred previously were identified: ellipsis and terminology. Firstly, ellipsis only accounted for 0.5% of the cases. We were not expecting this category to occur, since for the scope of customer support, all the content of an e-mail must be explicit, hence, the absence of important information obstructs the overall meaning. Secondly, as for terminology, we identified only one TM.

Thus, the following task was to analyze these context-dependent segments using a POS tagger to check for morpho-syntactic patterns, in order to verify if it was possible to automatically classify context-dependent TMs by using POS information. Ellipsis was the only category that was left out from this analysis since the element(s) that triggers contextual problems is omitted.

5.3. POS analysis in context-dependent vs. context-independent TMs

As mentioned in section 4.4.1, for this part of the experiment, we used Stanza, the POS tagger already used at Unbabel. The dataset consisted of a total of 670 TMs, in which 335 were context-dependent and 335 were context-independent. We had the same number of TMs for each type of segment, in order to have a balanced dataset. From the data, we excluded the TMs with ellipsis, since it is not possible to identify the POS category of the omitted elements.

5.3.1. POS tags distribution

Results show that the grammatical categories (see section 3.4) that triggered context related-issues were: verb (VERB), pronoun (PRON) and adjective (ADJ). *Figure 1* shows that the grammatical category with most occurrences is pronouns. The same POS categories mentioned above were also found in the context-independent segments (see *Figure 2*).

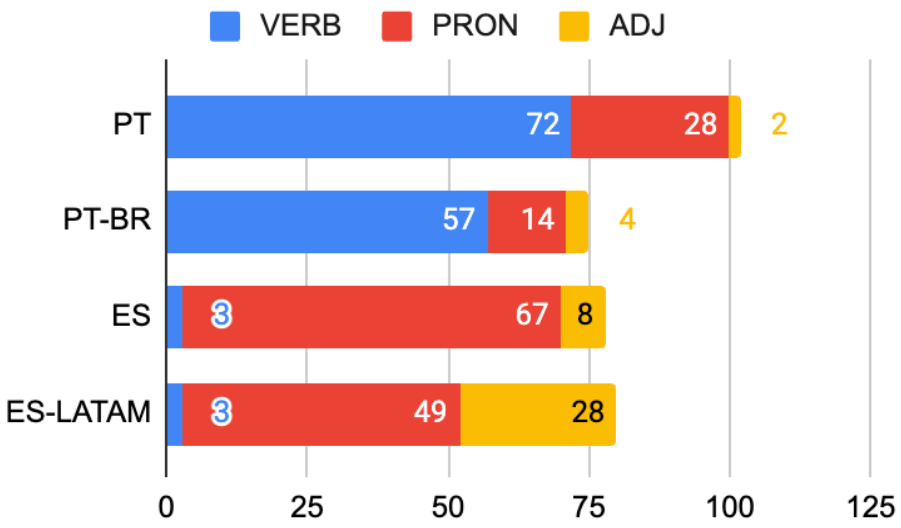

Figure 1 *- POS tags per LP for context-dependent TMs*

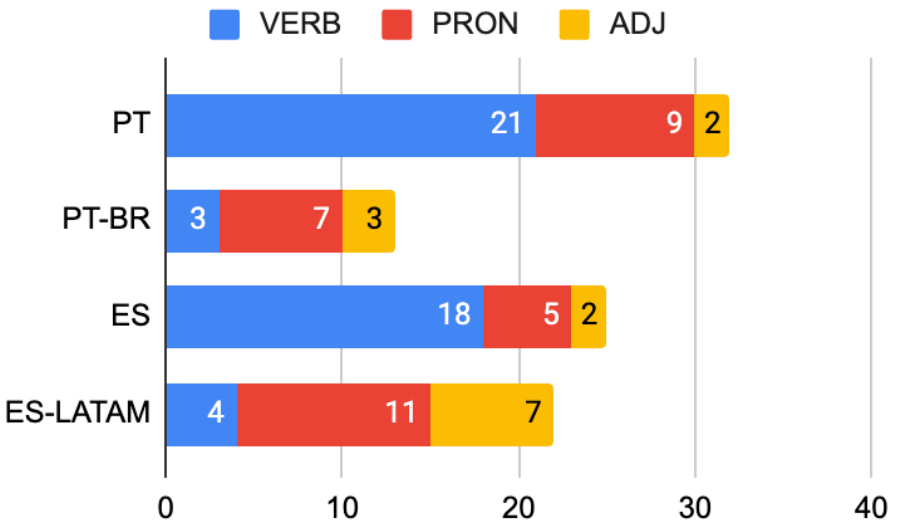

Figure 2 *- POS tags per LP for context-independent TMs*

When comparing the tags distribution, it is noticeable that even though the same tags occur in both types of segments they are much more frequent in the context-dependent segments than in the context-independent ones. For instance, pronoun was the category with most occurrences for all languages, totaling 47% of context-dependent cases, however, it only occurred 35% of the times in the context-independent segments. Verbs and adjectives were the most frequent categories for both types of segments. In total, for these segments, verbs totaled

50% of the cases, and adjectives 15%, while for the context-dependent they occurred in 40% and 13% of the segments, respectively.

An additional note should be stated. For both context-dependent and context-independent it is evident that these core categories are very relevant for both instances and with similar distributions. We were also expecting to observe nouns as a relevant category, but our initial intuition was not confirmed. The distribution of the core categories *per se* is not the main point we would like to emphasize. Our focus is on the pattern distribution of the parts of the segments that trigger context constraints.

The analyzed data is from the customer support domain, an interaction between an agent of a company and a client via ticket. This means that there are many grammatical categories such as pronouns and adjectives that refer to both entities. This also coincides with the fact that all the LPs are languages that are morphologically marked by grammatical gender. As for the verb category, the high percentage in both varieties of Portuguese is highly influenced by the use of participial forms of the verbs, such as "obrigado/obrigada". From a first inspection, we can say that the Portuguese varieties are highly influenced by participial forms, and the Spanish data is mostly classified as context-dependent due to the use of pronouns.

5.3.2. POS patterns distribution

Overall, from the previous analysis, namely the annotation of context annotation related issues in TMs, a total of 670 sentences were analyzed by the POS tagger system, 335 being context-dependent and 335, context-independent (see section 4.4.1). We now zoom in and check the detailed patterns. From this analysis, eight different patterns were found for the four language pairs: three only occurred in PT and PT-BR, one was unique for ES and ES-LATAM and the remaining four were common to all languages.

In the following sections, we will present the POS tags distribution per language pair found in this experience both in context-dependent and context-independent TMs. All the patterns gathered refer only to the sequence of words that causes contextual problems and not the entire sentence. The asterisks (*) represent the grammatical categories that trigger the context problems.

5.3.2.1. POS patterns for context-dependent and independent TMs for PT

For the Portuguese data, a total of 102 segments were annotated as context-dependent. In contrast, 102 context-independent segments were annotated as well, thus totaling 204 TMs (see *Table 10*). While all the context-dependent segments matched with one of the patterns, the same was not true for the context-independent ones, where only 35 (25%) sentences had a match.

	VERB+*PRON	*VERB+ADP	AUX+*VERB	AUX+*ADJ	*VERB	*PRON+VERB	Total:
Context-dependent	**25** (25%)	**54** (53%)	4 (4%)	2 (2%)	**14** (14%)	3 (3%)	102 (75%)
Context-independent	9 (26%)	17 (49%)	4 (11%)	2 (6%)	0	0	35 (25%)
Total:	34 (25%)	71 (52%)	8 (6%)	4 (3%)	14 (10%)	3 (2%)	137

***Table 10**. POS patterns subspecification for EN-PT*

As previously stated, for PT, the predominant grammatical category was "Verb". This category is related to the participial forms of the verb "obrigado/obrigada" (thank you, in English), commonly used as an expression of gratitude, which in Portuguese has a feminine and a masculine form and should be used accordingly. This resulted in two patterns: *VERB+ADP and *VERB.

The pattern *VERB+ADP corresponds to a sequence of a verb with gender constraints followed by a preposition (or adposition). It was the pattern with the highest number occurrences in the context-dependent segments, 53% of the data. It also had a fair number of occurrences in the context-independent, 49% (see examples 11 and 12). This pattern is common to both gender agreement and also Register (see example 11 below):

(11) Context-dependent:

EN: Thank you for contacting URL-0. [Informal register]

PT: [**Obrigado** por] [**entrar** em] contacto com a URL-0.

(12) Context-independent:

EN: I know you were expecting this trip to the original location.

PT: [Sei que] estava à espera desta viagem para o local original.

The pattern *VERB matched with 14% of the TMs and only in context-dependent data (example 13). Like the previous patterns shown, this one corresponds to TMs with only the verb "Obrigado(a)".

(13) Context-dependent:

EN: Thank you.

PT: **Obrigado**,

One other pattern with a significant number of context-dependent cases was VERB+*PRON (shown in the following example within brackets). It corresponds to a sequence of a verb followed by a pronoun, the pronoun being problematic usually because of gender agreement.

(14) Context-dependent:

EN**:** Therefore, we would like you to know that despite the news we have given, we can still help you find an alternative place and enjoy this trip without any difficulty.

PT**:** Portanto, gostaríamos que soubesse que, apesar das notícias que fornecemos, ainda podemos [ajudá-**lo**] a encontrar um local alternativo e a desfrutar desta viagem sem qualquer dificuldade.

(15) Context-independent:

EN: If you like this option, please follow the link above to make your new booking and send us your new confirmation number.

PT: Se esta opção for do seu agrado, por favor, efetue uma nova reserva na hiperligação acima e [envie-nos] o seu novo número de confirmação.

The AUX+*VERB pattern had the same number of occurrences for both context-dependent (4%) and independent (11%) segments. However, it was more frequent in the latter.

(16) Context-dependent:

EN: For more information on why it is being blocked, contact the card's issuing bank directly and:

PT: Para mais informações sobre a razão pela qual está a ser [**bloqueado**], entra diretamente em contacto com o banco emissor do cartão e:

(17) Context-independent:

EN: Your refund has been successfully processed.

PT: O seu reembolso [foi processado] com sucesso.

The last pattern found was *PRON+VERB, which is very similar to the previous one, changing only the order of the pronoun. It had no occurrences in the context-independent segments.

(18) Context-dependent:

EN: Rest assured that we will do our best in assisting you with this.

PT: Asseguro-lhe que faremos o possível para [**o**] ajudar.

A few conclusions can be drawn from PT patterns results. Firstly, *VERB and *PRON+VERB did not have occurances in the context-independent data. In addition, VERB+*PRON and *VERB+*ADP occurred for both context-dependent and context-independent segments, yet, the first pattern was more frequent among the context-independent (26%), while the latter was more frequent among the context-dependent (53%). The patterns, AUX+*VERB and AUX+*ADJ had the same number of occurrences for both cases. Even though they occur more often in context-independent than in context-dependent, they are not distinctive enough.

5.3.2.2. POS patterns for context-dependent and independent TMs for PT-BR

For Brazilian Portuguese, the dataset comprised a total of 150 segments, where 75 were context-dependent and 75 were context-independent. Similar to PT, six patterns were identified (see *Table 11*). While all context-dependent segments matched with one of the patterns, for the independent ones, there were only 26 segments that matched with a pattern, with one pattern not registering any match whatsoever.

	VERB+*PRON	*VERB+ADP	AUX+*VERB	AUX+*ADJ	VERB+*ADJ	*VERB	Total:
Context-dependent	**14** (17%)	**37** (49%)	4 (5%)	4 (2%)	2 (2%)	**14** (17%)	75 (74%)
Context-independent	5 (3%)	13 (3%)	5 (3%)	2 (1%)	1 (4%)	0	26 (26%)
Total:	19 (19%)	50 (50%)	9 (9%)	6 (4%)	2 (2%)	14 (15%)	101

Table 11. *EN-PT-BR POS patterns*

Similar to the PT, "Verb" was also the most frequent grammatical category for PT-BR and it was also related to the occurrence of the verb "Obrigado" (thank you). This category resulted in four patterns: *VERB, *VERB+ADP and AUX+*VERB.

The pattern *VERB+ADP (in 19), which corresponds to a sequence of a verb followed by a preposition, had most occurrences of context-dependent, matching 49% of the TMs, with only 3% for the context-independent.

(19) Context-dependent:

EN: Thank you for your response and for providing the information requested.

PT-BR: [**Obrigado**] pela sua resposta e por fornecer as informações solicitadas.

(20) Context-independent:

EN: Your PRS_ORG has been put in for deletion on the back end

PT-BR: Seu PRS_ORG foi [colocado para] exclusão no back-end.

The pattern *VERB, similar to what was registered for PT, only matched with context-dependent segments. It also represents segments with the instances of “Obrigado” as shown in the following example:

(21) Context-dependent:

EN: Thank you,

PT-BR: [**Obrigado**],

Other frequent pattern is the VERB+*PRON, which refers to a sequence of a verb followed by a problematic pronoun, in the sense that generates gender constraints and therefore requires context, as in example (22). This pattern was more frequent in context-dependent segments than in independent ones (17% vs. 3%, respectively).

(22) Context-dependent:

EN: You should be able to access it now.

PT-BR: Você deve ser capaz de acessá[**-lo**] agora.

(23) Context-independent:

EN: For future reference, your ticket number for this issue is PHONENUMBER-0 and you can reference it should you have any further questions regarding this case.

PT-BR: Para consultas futuras, o número do seu tíquete para esse problema é PHONENUMBER-0, e você pode [usá-lo] como referência se tiver mais dúvidas sobre esse caso.

The last pattern, whose sequence involved a verb, is AUX+*VERB. The results obtained between context-dependent and independent were not very disparate, but even so, the pattern had more instances for the latter (5% vs. 3%).

(24) Context-dependent:

EN: If you have been blocked from using PRS_ORG because the birthdate you entered upon sign up signifies that you are under 18 years old, you will remain blocked from the service for the amount of time specified on the login screen.

PT-BR: Se você foi [**impedido**] de usar o PRS_ORG porque a data de nascimento inserida ao se inscrever indica que você tem menos de 18 anos, você continuará [**bloqueado**] no serviço pelo período de tempo especificado na tela de login.

(25) Context-independent:

EN: I have checked our records and can confirm that your PIX payment has been processed and that the course has been added to your account.

PT-BR: Eu verifiquei nossos registros e posso confirmar que seu pagamento do PIX [foi processado] e que o curso foi adicionado à sua conta.

For all the patterns, the distinction between context-dependent and independent is due to the morphological properties of the grammatical categories that trigger gender constraints.

For this variant of Portuguese, *VERB was the only pattern that did not register any occurrences in context-independent TMs. We may state that it is a pattern that identifies context-dependent segments. The patterns VERB+*PRON and *VERB+ADP, even though occurred in context-independent segments, were more prominent in the context-dependent ones where they totaled 17% and 49% of these data. As for AUX+*ADJ and VERB+*ADJ, these were not very frequent in both segments, but even so, their occurrence was still higher in context-dependent than in context-independent TMs.

5.3.2.3. POS patterns for context-dependent and independent TMs for ES

For Spanish, the data comprised a total of 156 segments: 78 were annotated as context-dependent and 78 were context-independent. For this language, we identified five different patterns (see *Table 12*). Similar to the foregoing, only the context-dependent segments matched with one of the patterns. However, for the context-independent data, only 16 TMs matched with a pattern.

	ADP+*PRON	VERB+*ADJ	AUX+*ADJ	VERB+*PRON	AUX+*VERB	Total:
Context-dependent	**66** (85%)	5 (6%)	3 (4%)	1 (1%)	3 (2%)	78 (83%)
Context-independent	3 (3%)	1 (1%)	2 (2%)	4 (4%)	6 (6%)	16 (17%)
Total:	69 (74%)	6 (6%)	5 (5%)	5 (5%)	9 (9%)	94

Table 12. *POS patterns for ES*

Among all patterns, ADP+*PRON was the most frequent one, comprising 85% of the context-dependent segments, while only occurring in 3% of the context-independent ones. This pattern corresponds to a sequence of a proposition followed by a pronoun that has gender constraints. It also corresponds to the occurrences of the pronoun *nosotros* ("us", in english), which can have either a masculine or a feminine form (see example 26 and 27).

(26) Context-dependent:

EN: Thank you for getting in touch!

ES: ¡Gracias por ponerte en contacto con [**nosotros]**!

(27) Context-independent:

EN: Therefore, there is no entitlement to legal compensation in this case

PT: [Por lo] tanto, no hay derecho a una compensación legal en este caso.

Once again, it is the morphological properties of the pronoun in (26) and the context in which it occurs that distinguishes both sentences.

The pattern VERB+*ADJ refers to a sequence where there is a verb followed by an adjective that creates contextual problems due to gender constraints. It is more frequent for the context-dependent TMs (6%) than the context-independent ones (1%).

(28) Context-dependent:

EN: I will be happy to help you today.

ES: Estaré [**encantado**] de ayudarle hoy.

(29) Context-independent:

EN: Activation keys and DLC codes will remain linked to the deleted account and will not be usable on any other PRS_ORG account.

ES: Las claves de activación y los códigos DLC [permanecerán vinculados] a la cuenta eliminada y no podrán utilizarse en ninguna otra cuenta de PRS_ORG.

The pattern AUX+*ADJ refers to a sequence where there is a verb followed by an adjective that creates contextual problems. The results obtained were very similar, but even so, they were still higher in context-dependent segments (4% vs. 2%).

(30) Context-dependent:

EN: If you have any other questions, we are happy to help.

ES: Si tienes cualquier otra pregunta, estaremos [**encantados**] de ayudarte.

(31) Context-independent:

EN: Your PRS_ORG has been put in for deletion on the back end.

ES: Su PRS_ORG [está listo] para eliminarse en el back-end.

Lastly, the pattern AUX+*VERB matched with both gender agreement and also Register segments (see example 32). It had a higher occurrence in context-independent, occuring in 6% of the cases.

(32) EN: We hope that you continue to enjoy playing our game. [Formal register]

ES: Esperamos que [**sigas**] disfrutando de nuestro juego.

For this language pair, the pattern ADP+*PRON was the one that registered the highest number of instances among all patterns, accounting for 85% of all context-dependent cases. Similar to what was observed for PT-BR, the patterns VERB+*ADJ and AUX+*ADJ were not

as frequent as the previous pattern, however, frequency was higher in context-dependent TMs. As for the remaining two patterns, VERB+*PRON and AUX+*VERB, they were more frequent among the context-independent segments.

5.3.2.4. POS patterns for context-dependent and independent TMs for ES-LATAM

Lastly, for Latin-american Spanish, we found five different patterns. The dataset comprised 160 TMs, where 80 were context-dependent and 80 were context-independent (see *Table 13*). All the context-dependent segments matched with one pattern, however, not all context-independent segments had a match.

	ADP+*PRON	VERB+*ADJ	AUX+*ADJ	VERB+*PRON	AUX+*VERB	Total:
Context-dependent	**44** (55%)	**16** (20%)	**11** (14%)	6 (7%)	3 (4%)	80 (78%)
Context-independent	5 (5%)	3 (2%)	4 (2%)	6 (6%)	4 (4%)	22 (22%)
Total:	48 (48%)	19 (18%)	51 (10%)	12 (12%)	7 (7%)	102

***Table 13**. POS patterns for EN-ES-LATAM*

Similar to the results shown for ES, for this language, the pattern with most instances for context-dependent was ADP+*PRON, marking 55% of the context-dependent segments, contrasting with only 5% of occurrences in context-independent ones.

(33) Context-dependent:

EN: Please do not hesitate to reach out if you have any other questions.

ES-LATAM: Por favor, no dude en ponerse en contacto con [**nosotros**] si tiene alguna otra pregunta.

(34) Context-independent:

EN: They will follow up with you as soon as possible with a resolution.

ES-LATAM: Harán un seguimiento [con usted] lo antes posible con una resolución.

The pattern VERB+*ADJ (as in 35) was the second pattern with a higher number of occurrences in the context-dependent segments, totaling 20% vs. only 2% on the context-independent ones. In the example (36), the adjective in brackets *encantada* has its referent in the segment, therefore is context-independent. However, this was not the case for the adjective in (37), so the sentence is context-dependent.

(35) Context-dependent:

EN: You'll remain blocked from the service for the amount of time specified on the login screen.

ES-LATAM: Se mantendrá [**bloqueado**] desde el servicio por el tiempo especificado en la pantalla de inicio de sesión.

(36) Context-independent:

EN: Hi Camila, my name is Valentina with PRS_ORG support and I'll be happy to assist you with your Name Correction request.

ES-LATAM: Buen día, Camila, mi nombre es Valentina del soporte PRS_ORG y [estaré encantada] de ayudarle con su solicitud de corrección de nombre.

The pattern AUX+*ADJ had a similar behavior to the previous ones, occurring more frequently in context-dependent segments (14%) than in context-independent ones (2%).

(37) Context-dependent:

EN: We have received your request and are excited to assist you.

ES-LATAM: Hemos recibido su solicitud y estamos [**emocionados**] de ayudarle.

(38) Context-independent:

EN: I am looking forward to helping you, please standby and I'll be with you as soon as possible!

ES-LATAM: Espero poder ayudarle, por favor, espere y le contactaremos tan pronto como [sea posible].

As for VERB+*PRON (exemplified in 39), it had the same number of occurrences for both context-dependent and context-dependent, six segments for each (7% vs. 6%, respectively).

(39) Context-dependent:

EN: We look forward to helping you again in the future.

ES-LATAM: Esperamos ayudar[**lo**] nuevamente en el futuro.

(40) Context-independent:

EN: Thanks for reaching out and we apologize for any confusion.

ES-LATAM: Gracias por [contactarnos] y nos disculpamos por cualquier confusión.

The difference between the examples above relies on the morphological properties of the anaphoric pronouns. While in (39), the pronoun inflects in gender, meaning that has both a masculine and feminine form according to the context that it occurs, the pronoun in (40) is neutral, therefore, is context-dependent.

Lastly, the AUX+*VERB pattern had a similar frequency (4%) in both types of segments, hence it is not contrastive enough in the present dataset.

(41) Context-dependent:

EN: If you have been blocked from using PRS_ORG because the birthdate you entered upon sign up signifies that you are under 18 years old, you will remain blocked from the service for the amount of time specified on the login screen.

ES-LATAM: Si se le ha bloqueado el uso de PRS_ORG debido a que la fecha de nacimiento que ingresó al momento de registrarse significa que usted tiene menos de 18 años, permanecerá [**bloqueado**] del servicio por el tiempo especificado en la pantalla de inicio de sesión.

(42) Context-independent:

EN: Just confirming that we received your request and I've forwarded it over to the best team to handle it.

ES-LATAM: Solo [estoy confirmando] que recibimos su solicitud y la envié al mejor equipo para manejarla.

The results for ES-LATAM were slightly different from those obtained for ES. Firstly, the pattern ADP+*PRON obtained the most matches, and is, therefore, a context-dependent pattern identifier in this dataset. The patterns VERB+*ADJ and AUX+*ADJ were also very frequent among the context-dependent segments, occurring in 20% and 14% of these segments. The pattern VERB+*PRON had the same instances for both types of segments. Lastly, the pattern AUX+*VERB had a similar number of occurrences in both types of segments.

Overall, from the current analysis, eight POS patterns were gathered for context-dependent TMs: three that were exclusive for PT and PT-BR, namely *VERB+ADP, *VERB and *PRON+VERB, one that was also exclusive for ES and ES-LATAM, ADP+*PRON, and lastly, four patterns that were common among the four languages: VERB+*ADJ, AUX+*ADJ, AUX+*VERB and VERB+*PRON (Table 14 shows the distribution of these patterns per language).

As for the pattern distribution, it is important to mention that most of the patterns occurred in both context-dependent and independent segments. However, the number of matches was mainly higher among the context-dependent data. For PT and PT-BR, the pattern *VERB only matched with context-dependent segments. For these LPs, the pattern *VERB+ADP was also very frequent among the context-dependent segments. As for ES and ES-LATAM, the occurrence of the pattern ADP+*PRON was also higher among the TMs that needed context. The patterns VERB+*ADJ, AUX+*ADJ and VERB+*PRON were also very frequent for PT and PT-BR but not for ES and ES-LATAM. The only pattern that was not distinctive enough between context-independent and context-dependent was AUX+*VERB for all the four languages.

Languages	POS Patterns	Context-dependent	Context-independent
	*VERB	✅	

PT PT-BR	*VERB+ADP	✅	
	*PRON+VERB	✅	
ES ES-LATAM	ADP+*PRON	✅	
PT PT-BR ES ES-LATAM	VERB+*ADJ	✅	
	AUX+*ADJ	✅	
	VERB+*PRON	✅	✅
	AUX+*VERB	✅	✅

Table 14 *- Patterns distribution*

The analyzed content is from customer support tickets which represent interactions between an agent and a customer. This also coincides with the fact that all the LPs are languages that are morphologically marked by grammatical gender. Many of these categories occur in the 3rd person singular, with the forms “he” or “she”. Nevertheless, we were expecting that nouns would also appear within the aforesaid categories. This grammatical category is very frequent and allows us to refer to an addressee. However, it did not occur in any segment.

5.4. Validation of the POS patterns in context-dependent vs. context-independent TMs

From the preceding analysis, eight POS patterns were gathered: three exclusive for Portuguese and Brazilian Portuguese and one exclusive for Spanish and Latin-american Spanish and four that were common for all. The goal of current analysis was to verify if it would be possible to automatically classify segments as to their context, that is, whether or not it would be possible to use the POS information to identify context-dependent segments. Therefore, we searched for the POS patterns described above in a dataset of 8,000 TMs (2,000 TMs *per* language pair) from seven company’s clients from different domains that were never analyzed before.

Several results were obtained from this experiment. In total, 7,148 segments were context-independent. However, only 48% of these TMs matched with a pattern (see *Table 15*).

Context-independent			
	Matched with a pattern	Did not match with a pattern	Total:
PT	1,034 (54%)	877 (46%)	1,911 (27%)
PT-BR	749 (45%)	961 (55%)	1,710 (24%)
ES	780 (43%)	1,044 (57%)	1,824 (25%)
ES-LATAM	844 (51%)	859 (49%)	1,703 (24%)
Total:	3,407 (48%)	3,741 (52%)	7,148

***Table 15**. Results from context-independent*

As for the context-dependent segments, 75% of TMs matched with a pattern, and 25% did not fit into one of the patterns found. These ones were then annotated and grouped into new patterns (see *Table 16*).

Context-dependent			
	Matched with a pattern	Did not match with a pattern	Total:
PT	88 (99%)	1 (1%)	89 (11%)
PT-BR	190 (66%)	100 (34%)	290 (35%)
ES	171 (97%)	5 (3%)	176 (22%)
ES-LATAM	160 (61%)	102 (39%)	262 (32%)
Total:	609 (75%)	208 (25%)	817

***Table 16**. Results from context-dependent*

Overall, out of the 8,000 analyzed segments, only 4,224 TMs matched with one of the patterns found for both context-dependent and context-independent (see *Table 17*). The results showed that 81% of the segments were context-independent and 19% were context-dependent.

Needs Context	PT	PT-BR	ES	ES-LATAM	Total:
False	1,034 (92%)	749 (73%)	780 (82%)	844 (77%)	3,407 (81%)
True	90 (8%)	290 (27%)	176 (18%)	262 (23%)	818 (19%)
Total:	1,124 (26%)	1,039 (24%)	956 (22%)	1,106 (27%)	4.224

Table 17*. Results from experiment 3*

5.4.1. POS tags distribution

In the previous analysis (see section 5.3), we considered all the TMs that were annotated with gender and register issues. However, for this analysis, we only considered segments with gender agreement, due to the fact that this is the most frequent category found in all of the analyzed languages.

As in the previous experiment, VERB, PRON and ADJ were still the most prevalent POS categories in the analysis. However, two new categories were found: nouns (NOUN) and determiners (DET). *Table 18* shows the distribution of tags *per* language pairs for the context-dependent TMs with gender constraints.

	VERB	PRON	ADJ	NOUN	DET	Total:
PT	61 (67%)	15 (17%)	13 (15%)	1 (1%)	-	90 (11%)
PT-BR	72 (25%)	70 (24%)	76 (25%)	33 (12%)	39 (14%)	290 (35%)
ES	2 (1%)	126 (72%)	48 (27%)	-	-	176 (22%)
ES-LATAM	-	69 (26%)	103 (39%)	87 (33%)	3 (1%)	262 (32%)
Total:	135 (16%)	280 (34%)	240 (29%)	121 (15%)	42 (5%)	818

Table 18*. POS categories per LP for context-dependent*

As for the context-independent patterns distribution, the results *per* category were substantially higher with comparison to the context-dependent ones. However, the categories NOUN and DET did not have a match with a TM. *Table 19* shows the distribution of tags *per* language pairs for the context-independent data.

	VERB	PRON	ADJ	NOUN	DET	Total:
PT	688 (66%)	338 (33%)	8 (1%)	-	-	1,034 (30%)
PT-BR	454 (59%)	178 (23%)	117 (15%)	-	-	749 (22%)
ES	387 (49%)	302 (39%)	91 (12%)	-	-	780 (23%)
ES-LATAM	325 (42%)	325 (42%)	194 (23%)	-	-	844 (25%)
Total:	1,854 (54%)	1,143 (34%)	410 (12%)	-	-	3,407

Table 19. *POS categories per LP for context-independent*

The results obtained for the current experiment per category were not consistent with those of the previous analysis (see section 5.3.1.). Whereas previously there were a total of 331 tags for context-dependent and only 92 for context-independent, for this analysis, the opposite was verified, with only 817 for the former and 3.442 for the latter.

5.4.2. POS patterns distribution

In addition to the eight patterns already found in the previous experiment, we found seven new patterns in this analysis for the context-dependent TMs. In the following sections, we will present the POS patterns distribution per language pair. For all languages, we will first analyze the patterns that were previously identified as related with context-dependent TMs, and then, we will perform the same analysis on the new patterns found in this enlarged dataset.

Note that only the context-dependent segments that present gender constraints, where it was not possible to assign gender due to lack of context evidence, will be analyzed.

5.4.2.1. POS patterns for PT

For PT, a total of 1,124 TMs matched with one of the seven patterns. The results showed that a total of 1,034 segments were context-independent and 90 were context-dependent (see *Table 20*). A new pattern that only covered one context-dependent TM was also registered. As for patterns distribution, *PRON+VERB did not match with a context-dependent TM; and the

pattern *VERB only occurred in context-dependent segments that did not match with a context-independent, therefore, being consistent with the previous analysis (see section 5.3.2.1).

	VERB+* PRON	*VERB+ ADP	AUX+* VERB	AUX+* ADJ	*VERB	*PRON+ VERB	New patterns	Total:
Context-dependent	**15** (17%)	**54** (61%)	3 (3%)	**13** (15%)	4 (4%)	0	1 (1%)	90 (8%)
Context-independent	198 (19%)	535 (52%)	153 (14%)	8 (1%)	0	140 (13%)	0	1,034 (92%)
Total:	213 (18.9%)	589 (52.4%)	156 (13.8%)	21 (1.8%)	4 (0.4%)	140 (12.5%)	1	1,124

***Table 20**. POS patterns distribution for PT*

As in the previous analysis, all the patterns involving verbs, namely *VERB+ADP and *VERB, were related to the word “Obrigado” or “Obrigada” and were very frequent.

The pattern *VERB+ADP was the most recurrent among the context-dependent segments, corresponding to 61% of the cases. However, this pattern also occurred frequently in context-independent segments (52%). The distinction between context-independent and context-dependent segments is the morphological features. In the example shown in (43) the participial verb is followed by the preposition *pela* which is contracted due to gender variations. These cases would be very interesting to analyze in a future experiment.

(43) Context-dependent:

EN: Thanks for your reply.

PT: [**Obrigado** pela] sua resposta.

(44) Context-independent:

EN: Please let me know if you need anything else.

PT: Por favor, contacte-me se [necessitar de] ajuda em mais alguma coisa!

The pattern *VERB has no occurrences in context-independent segments, therefore, only occurring in context-dependent ones. This type of pattern usually corresponds to a closing segment with only the word “Obrigado” as in (45):

(45) EN: Thanks,
PT: [**Obrigado**],

In Portuguese, the expression mentioned above is both a feminine and masculine form that needs to be used according to the context, hence its classification as context-dependent.

As in the previous analysis, the pattern AUX+*VERB remained prevalent within the context-independent segments, marking 14% of these cases. Similar to the verb “Obrigado”, the verb in (46) also is in its participial form:

(46) Context-dependent:
EN: You’ll be notified right away if your photo was accepted or denied
PT: [Será **informado**] imediatamente quando a sua fotografia for aceite ou recusada.

(47) Context-independent:
EN: We’re sorry to hear about the trouble you’re experiencing.
PT: [Lamentamos saber] que está a ter problemas.

For this analysis, and similar to the results obtained for this language previously, the pattern AUX+*ADJ registered more context-dependent TMs (15%) than context-independent ones (1%). This category is shown in examples (48) and (49):

(48) Context-dependent:
EN: I’m glad to hear that you’re interested in verifying your profile.
PT: Fico feliz por saber que está [**interessado**] em verificar o seu Perfil.

(49) Context-independent:

EN: We were not able to find a matching profile based on the information you provided.

PT: Não nos [foi possível] encontrar um perfil correspondente, com base nas informações que nos forneceu.

Unlike the previous experiment, the pattern *PRON+VERB did not match with a context-dependent segment.

During the analysis, we registered one new pattern for a context-dependent TM: *DET+*NOUN, corresponding to a sequence of a determiner and a noun, as it shows in (50):

(50) Source: A checkmark will appear on your profile, signaling to potential matches that you're really **you.**

Target: Vai aparecer uma marca de verificação no seu Perfil, a indicar a potenciais matches que realmente é [**o senho**r].

Similar to the results obtained from the previous analysis for this language pair, the pattern *VERB remained the only one that occurred in context-dependent cases, since it has no occurrences in the context-independent TMs. Also, AUX+*ADJ was more frequent in context-dependent segments, than in the context-independent ones. The pattern *VERB+ADP, was also very frequent among the context-dependent cases (61%). Unlike these, VERB+*PRON and AUX+*VERB, although matching with a considerable number of segments that needed context, were more frequent in context-independent data. In addition, to the previous patterns a new was registered: *DET+*NOUN.

5.4.2.2. POS patterns for PT-BR

For the PT-BR data, as far as results are concerned, a total of 1,062 TMs matched with a pattern, in which 773 segments were context-independent therefore comprehending 72% of the total dataset; in contrast, 290 segments (28%) were context-dependent. Out of these, 100 corresponded to new patterns. The pattern *VERB did not match with a context-independent segment, therefore being in accordance with the previous analysis (see section 5.3.2.2.)

Nonetheless all other patterns had a match (see *Table 21*). In addition to this, seven new patterns were found for the context-dependent data.

	VERB+*PRON	*VERB+ADP	AUX+*VERB	AUX+*ADJ	VERB+*ADJ	*VERB	Total:
Context-dependent	**68** (36%)	**65** (34%)	2 (1%)	**47** (25%)	6 (3%)	2 (1%)	190 (20%)
Context-independent	178 (23%)	318 (41%)	136 (13%)	117 (15%)	24 (3%)	0	773 (80%)
Total:	246 (25.5%)	383 (39.8%)	138 (14.3%)	164 (17.0%)	30 (3.1%)	2 (0.2%)	963

***Table 21**. POS patterns distribution for PT-BR*

The pattern VERB+*PRON had the highest occurrences among the context-dependent segments, corresponding to 36% of the cases. However, this pattern was not distinguishable enough from the context-independent ones, with 23% of occurrences..

(51) Context-dependent:

EN: I'm extremely glad to be assisting you today.

PT-BR: Fico extremamente feliz em [ajudá-**la**] hoje.

(52) Context-independent:

EN: I'd love to see you complete several missions and earn more amazing rewards!

PT-BR: Eu adoraria [ver você] concluir várias missões e ganhar mais recompensas incríveis!

For PT-BR, the pattern *VERB+ADP (as in 53 and 54) occurred in 34% of the context dependent TMs and 41% in the context-independent ones. Instances of this pattern are illustrated in the following examples:

(53) Context-dependent:

EN: Thanks for contacting the PRS_ORG Support Team.

PT-BR: [**Obrigada** por] entrar em contato com a Equipe PRS_ORG.

(54) Context-independent:

EN: Hi there, we appreciate your time in leaving us this review.

PT-BR: Olá, [agradecemos pelo] comentário enviado!

For both of the two last patterns, the distinguishing factor between context-independent from context-dependent is the verb type morphological features of the selected verbs. While in (53), the verb is a participial verb and inflects in gender, this is, can have both a masculine and feminine form, the one in (54) is neutral and therefore can occur without gender constraints.

The pattern AUX+*VERB (examples 55 and 56) only matched with two context-dependent segments, being more frequent in context-independent TMs (1% vs. 13%, respectively).

(55) Context-dependent:

EN: I know you are going to conquer, knowing how determined of a PRS_ORG you are!

PT-BR: Eu sei que você vai ganhar, sabendo o quanto você [é **determinada**] com o PRS_ORG!

(56) Context-independent:

EN: I'm Silva, and I will be assisting you today.

PT-BR: Eu sou Silva e [estarei ajudando] você hoje.

As for, VERB+*ADJ the results were also higher for context-independent TMs, corresponding to 3% of these cases but also 3% for the context dependent ones (examples 57 and 58):

(57) Context-dependent:

EN: If you were satisfied with my assistance, please take the survey when you get one.

PT-BR: Se você ficou [**satisfeito**] com a minha assistência, por favor, responda à pesquisa quando você receber uma.

(58) Context-independent:

EN: Happy to e-meet you and it's my pleasure to assist you.

PT-BR: [Fico feliz] em te atender e é um prazer te ajudar.

As for the pattern AUX+*ADJ, it also had more instances among the context-independent data, occuring in 15% of these cases and 25% of the context-dependent ones. This pattern is exemplified following (see example 59 and 60):

(59) Context-dependent:

EN: Stay safe & take care

PT-BR: Fique [**segura**] e cuide-se,

(60) Context-independent:

Source: You were phenomenal in the game.

Target: Você [foi fenomenal] no jogo.

Lastly, the pattern *VERB corresponds to the instances of the verb “Obrigado/Obrigada” as shown in the example (61). Similar to PT, this pattern only occurred among the context-dependent TMs.

(61) EN: EMOJI-0 Thank you!

PT-BR: EMOJI-0 [**Obrigada**]!

5.4.2.2.1. New POS patterns

For PT-BR, we found seven new patterns for the context-dependent TMs (see *Table 22*). The domain analyzed was gaming, generally marked by informal language and specific jargon

that involves nouns, hence the higher number of nouns and determiners than in the previous language pairs.

	*DET+*NOUN	*DET+*ADJ+*NOUN	DET+NOUN+*ADJ	ADP+*NOUN	*PRON+AUX+*VERB	*VERB+CCONJ	AUX+*NOUN	Total:
Context-dependent	40 (14%)	7 (2%)	18 (6%)	26 (9%)	3 (1%)	2 (1%)	3 (1%)	100 (10%)

Table 22. *New POS patterns for PT-BR*

The patterns *DET+*NOUN (example 62), *DET+*NOUN+*ADJ (example 63) and *DET+*ADJ+*NOUN (example 64) are very similar, changing only the position of the adjective, which may or may not occur at all in the sequence.

(62) EN: As a player myself, I too would be sad to miss out on any rewards or progress in the game

PT-BR: Sendo [uma **jogadora**] também, eu também ficaria triste em perder quaisquer recompensas ou progresso no jogo.

(63) EN: I'm João, and I'm delighted to assist a fantastic player like you.

PT-BR: Eu sou João e estou muito feliz em ajudar [**uma jogadora fantástica**] como você.

(64) EN: I appreciate you being our valuable player and for progressing so far in the game.

PT-BR: Agradeço por ser [**o nosso valioso jogador**] e por progredir até agora no jogo.

The nouns in bold correspond to *jogadora* or *jogador* (meaning “player” in English) and have gender constraints.

The pattern ADP+*NOUN refers to a sequence of an adposition and a noun, the latter being problematic again because the noun may vary in gender.

(65) EN: As a player, I can understand how bothersome that would be.
PT-BR: [Como **jogador**], posso entender como isso seria incômodo.

The pattern *PRON+AUX+*VERB is also a variation of the pattern AUX+*VERB already found in the previous analysis. However, this one involves a pronoun.

(66) EN: They have been properly added to your account without any error.
PT-BR: [**Elas** foram **adicionadas**] corretamente à sua conta sem nenhum erro.

The pattern *VERB+CCONJ corresponds to a verb followed by a conjunction as in (67). Compared to the previous patterns, this one occurred less often (only in 1% of the cases).

(67) EN: Thank you and have a good day!
PT-BR: [**Obrigada**] e tenha um bom dia!

Lastly, AUX+*NOUN refers to a sequence of a verb and a noun. This pattern only matched with one context-dependent segment.

(68) Context-dependent:
EN: It is players like you who keep our community growing rapidly.
PT-BR: [São **jogadores**] como você que mantêm nossa comunidade crescendo rapidamente.

The results for the current analysis for PT-BR showed that the pattern *VERB continued to only identify segments that need context, thus not occurring in the context-independent ones. This was aligned with the previous results for this language (see section 5.3.2.1.). On the contrary, the patterns AUX+*VERB and VERB*ADJ had a low percentage of occurrence among the context-dependent segments where it scored 1% and 6%, respectively, thus presenting a

considerable number of context-independent segments. The patterns *VERB+ADP, AUX+*ADJ and VERB*PRON also occurred in a high number of context-independent segments. However, the percentages obtained by the patterns show that these were more frequent in context-dependent TMs.

For this experience, seven new patterns were found. These new patterns mostly involved nouns, a POS category that did not occur in the previous experiment. We may assume that the emergence of these new patterns, with mainly the NOUN category, can be related with the gaming domain.

5.4.2.3. POS patterns for ES

For Spanish, the results showed that a total of 956 TMs matched with one of the patterns found, in which 780 (82%) were context-independent and 176 (18%) were context-dependent. In addition to the five patterns, six new ones were found (see *Table 23*). As for patterns distribution, apart from the AUX+*VERB, which had no occurrences in context-dependent TMs, there were matches for all patterns. In addition to these, we also found three new patterns for context-dependent segments, totalling eight patterns for ES.

	ADP+*PRON	VERB+*ADJ	AUX+*ADJ	VERB+*PRON	AUX+*VERB	New patterns	Total:
Context-dependent	**116** (66%)	**15** (9%)	**31** (18%)	9 (5%)	0	5 (2%)	176 (18%)
Context-independent	67 (9%)	25 (3%)	70 (9%)	235 (30%)	380 (49%)	0	780 (82%)
Total:	183 (19%)	40 (4%)	101 (10%)	244 (26%)	380 (40%)	5 (1%)	956

***Table 23**. POS patterns distribution for ES*

The pattern ADP+*PRON was the most frequent context-dependent pattern, corresponding to 66% of the data. In the context-independent segments, it only occurred in 9% of the cases. For this pattern, the distinction between both types of segments was the morphological properties of the pronouns as shown in (69) and in (70):

(69) Context-dependent:

EN: If there is anything else we could help you with, please do not hesitate to contact us again.

ES: Si hay algo más en lo que te podamos ayudar, por favor, no dudes en ponerte en contacto [con **nosotros**] de nuevo.

(70) Context-independent:

EN: We look forward to speaking with you soon with an update.

ES: Esperamos hablar [con usted] pronto con una actualización.

The pattern VERB+*PRON matched 30% of the context-independent segments and only 5% of the context-dependent ones. Examples of this pattern are shown in (71) and in (72):

(71) Context-dependent:

EN: If so, could you try to disable them?

ES: Si es así, ¿puedes probar a [desactivar**los**]?

(72) Context-independent:

EN: Thank you for contacting us.

ES: Gracias por [contactarnos].

The pattern VERB+*ADJ had very similar results between both types of segments, however, it was more frequent in the context-dependent, totaling 9% of the cases, than in the context-independent ones, totaling 3% (see examples 73 and 74):

(73) Context-dependent:

EN: Please let me know if you need anything else - I will be more than happy to help.

ES: Por favor, hágame saber si necesita algo más, [estaré **encantado**] de ayudarle.

(74) Context-independent:

EN: I am so sorry to hear that your order has arrived incomplete and I would like to resolve this for you as quickly as possible.

ES: Lamento mucho saber que su pedido ha [llegado incompleto] y me gustaría resolver esto tan pronto como sea posible.

The pattern AUX+*ADJ is similar to the previous one, only changing the main category of the verb. The results showed that this pattern was more frequent among the context-dependent TMs (18%) than in the context-independent ones (9%).

(75) Context-dependent:

EN: It's great to hear that you're interested in our products!

ES: ¡Es genial saber que [está **interesado**] en nuestros productos!

(76) Context-independent:

EN: Thank you for your email, I am happy your product arrived!

ES: Gracias por su correo electrónico, ¡[estoy feliz] de que su producto haya llegado!

Lastly, the pattern AUX+*VERB did not match with a context-dependent segment. It matched, however, with 387 context-independent segments, corresponding to almost half of these TMs (40%). An example is shown in (77):

(77) Context-dependent:

EN: This has to be a new document, and can't be something you've already submitted.

ES: Tiene que ser un nuevo documento, no algo que ya [haya enviado].

5.4.2.3.1. New patterns for ES

For ES, we found three new patterns that matched with only five context-dependent segments.

The first pattern, ADV+*ADJ corresponds to the sequence of an adverb (ADV) and a problematic adjective (ADJ). It matched with two TMs (example 78):

(78) EN: Your feedback is so important to me, and if you have a spare moment, I would be so grateful if you could leave me a review via our PRS_ORG page: URL-0.

ES: Sus comentarios son muy importantes para mí, y si tiene un momento libre, estaré [muy **agradecido**] si pudiera dejarme una revisión a través de nuestra página de PRS_ORG: URL-0.

The pattern *VERB+ADP had already occurred in other languages, PT and PT-BR respectively, however, it occurred for the first time in the present ES dataset, for one TM. This pattern is shown in example (79):

(79) EN: Please send from a Post Office in order to obtain a proof of postage receipt and retain for your records

ES: [**Envíelo**] desde una oficina de correos para obtener un comprobante de recibo postal y guárdelo para sus registros.

The pattern *PRON+AUX corresponds to a pronoun followed by a auxiliary verb as shown in the following example:

(80) EN: They will be able to recommend some products that you will love!

ES: ¡[**Ellos**] podrán recomendarle algunos productos que le encantarán!

For ES, ADP+*PRON remains the pattern whose instances were always higher in context-dependent segments, marking 66% of these cases. In contrast, it only occurred 9% of the time in context-independent TMs. These results align with those obtained in the previous analysis (see section 5.3.2.3). As for VERB+*ADJ and AUX+*ADJ the occurrences were always higher in the context-dependent segments. In contrast, for VERB*PRON, the patterns

were very frequent for the context-independent. Lastly, AUX+*VERB was the only pattern that did not match with a context-dependent segment, therefore, it remained as a pattern that is not distinctive between context-dependent and context-independent TMs. Nevertheless, three new patterns were found during the analysis, even though with low frequency (2%).

5.4.2.4. POS patterns for ES-LATAM

Lastly, for ES-LATAM, a total of 1,106 TMs matched with a pattern: 844 (76%) were context-independent and 262 (14%) were context-dependent. In addition to the five patterns found previously for this language pair, we also found seven new patterns in this dataset. Except for the pattern AUX+*VERB, that did not match with context-dependent TMs, all other segments matched with a pattern (see *Table 24*).

	ADP+*PRON	VERB+*ADJ	AUX+*ADJ	VERB+*PRON	AUX+*VERB	Total:
Context-dependent	**58** (36%)	10 (6%)	**81** (51%)	11 (7%)	0	160 (14%)
Context-independent	58 (7%)	16 (2%)	143 (17%)	302 (36%)	325 (42%)	844 (76%)
Total:	116 (12%)	26 (2%)	224 (22%)	313 (31%)	325 (32%)	1,004

***Table 24**. POS patterns for EN-ES-LATAM*

The pattern ADP+*PRON was much more frequent in context-dependent segments (36%) than in independent ones (7%)(see examples 81 and 82).

(81) Context-dependent:

EN: If you require any more assistance, feel free to contact us.

ES-LATAM: Si necesitas más ayuda, no dudes en ponerte en contacto con [**nosotros**].

(82) Context-independent:

EN: My assistance for you doesn't end here, feel free to write to us for any game-related queries/suggestions.

ES-LATAM: Mi ayuda [para ti] no termina aquí, no dudes en escribirnos para cualquier consulta o sugerencia relacionada con el juego.

In both examples, the main difference between both segments is the morphological features. While the pronoun *nosotros* in (81) can have either a masculine and feminine form, the clitic pronoun *ti* (82) is neutral.

The pattern AUX+*ADJ was the pattern with most instances for the context-dependent TMs, totaling 51% of the cases, where it only marked 17% of the context-independent data (example in 83).

(83) Context-dependent:

(a) EN: I'll be happy to assist you today.

ES-LATAM: Estaré [**encantado**] de ayudarte hoy.

(84) Context-independent:

(a) EN: It's great to hear from you again.

ES-LATAM: [Es genial] saber de ti de nuevo.

The pattern VERB+*ADJ was the one with fewer cases in both context-dependent and context-independent, however, it was more frequent in the former, totaling 6% and only 2% in the latter.

(85) Context-dependent:

(a) EN: We will be happy to help you.

ES-LATAM: Estaremos [**encantados**] de ayudarte.

(86) Context-independent:

(a) EN: Please make sure that you have enough storage space on your device and then try again.

ES-LATAM: Por favor, asegúrate de [tener suficiente] espacio de almacenamiento en tu dispositivo y luego inténtalo de nuevo.

Regarding the pattern VERB+*PRON, it is less frequent in context-dependent segments (7%) than in context-independent ones, occurring in 36% of the cases (examples 87 and 88).

(87) Context-dependent:

(a) EN: Please restart your game to collect it.

ES-LATAM: Por favor, reinicia tu juego para obtener[**la**].

(88) Context-independent:

(a) EN: Happy to e-meet you and it's my pleasure to assist you.

ES-LATAM: Me alegra [conocerte] y es un placer ayudarte.

Similarly to ES, the patterns AUX+*VERB (as in 89) did not match with a context-dependent TM in the dataset, but was frequent among the context-independent cases (42%).

(89) EN: Then restart the game and you'll be able to collect your reward.

ES-LATAM: A continuación, reinicia el juego y [podrás obtener] tu recompensa.

5.4.2.4.1. New patterns

Similar to the other LPs mentioned above, for ES-LATAM we found seven new patterns (see *Table 25*), that totaled 102 context-dependent segments.

	SCONJ+*NOUN	*DET+*NOUN	*DET+*NOUN+*ADJ	*NOUN+*ADJ	*DET+*PROPN	ADV+*ADJ	AUX+*NOUN	Total:
Context-dependent	32 (3%)	23 (2%)	30 (3%)	5 (0.4%)	3 (0.3%)	8 (1%)	1 (0.1%)	102 (9%)

***Table 25**. New POS patterns for EN-ES-LATAM*

The pattern SCONJ+*NOUN corresponds to a sequence of a subordinating conjunction followed by a noun as it in (90):

(90) EN: As a player myself, I too would be sad to miss out on any rewards or progress in the game.

ES-LATAM: [Como **jugador**], también me entristece perder las recompensas o el progreso en el juego.

The patterns *DET+*NOUN, *DET+*NOUN+*ADJ and *NOUN+*ADJ are similar, determiners and adjectives may be optional. Examples (90) to (96) show this:

(91) EN: Being a player myself I too would be unhappy if I see my efforts not paying out.

ES-LATAM: Al ser [**un jugador**], también me sentiría triste si viera que mis esfuerzos no están siendo pagados.

(92) EN: I would like to thank you for being a valued player.

ES-LATAM: Me gustaría agradecerte por ser [**un jugador valioso**].

(93) EN: I appreciate you being our valued player and for progressing so far in the game.

ES-LATAM: Agradezco que seas nuestro [**jugador valioso**] y por progresar hasta ahora en el juego.

The pattern *DET+PROPN corresponds to a sequence of a determiner and a neutral proper name.

(94) EN: Being a fellow PRS_ORG, I know how that would be for you.

ES-LATAM: Al ser [**un** PRS_ORG], sé cómo sería para ti.

The pattern AUX+*NOUN refers to a sequence of an auxiliary verb followed by a name as shown in the example (95):

(95) EN: Being a player myself, I understand how upsetting this must be for you.
ES-LATAM: También soy [**jugador**] y entiendo lo molesto que esto debe ser para ti.

The pattern AUX(+ADV)+*ADJ is a variation of the pattern AUX+ADJ mentioned above, however with one adverb occurring between the two POS categories.

(96) EN: I'm so glad you contacted me about that!
ES-LATAM: ¡[Estoy muy **contento**] de que me hayas contactado sobre eso!

For this language variant, similarly to the results obtained previously, the pattern ADP+*PRON was more frequent in the context-dependent TMs. Also, similar to the findings for ES (shown in the previous section) the pattern AUX+*VERB did not match with a segment that needed context. The patterns, VERB+*ADJ and AUX+*ADJ and more frequent in context-dependent TMs than in context-dependent. As for VERB+*PRON, it was more frequent for context-independent TMs. For this analysis, seven new patterns were found for this analysis, totaling 9% of the context-dependent segments. Similar to the PT-BR (see section 5.4.2.2.), most of these new patterns involved nouns. Once again, this could be related to the gaming domain.

Overall, most of the segments found were context-independent, marking 76% of these cases, in contrast, only 24% of them were context-dependent. As for the POS distribution, the patterns *VERB for PT and PT-BR and ADP+*PRON, for ES and ES-LATAM, only occurred in context-dependent segments. This was consistent with the previous experiment, therefore, we can conclude that these patterns can identify context-dependent segments. The remaining patterns, *VERB+ADP, VERB+*ADJ, AUX+*ADJ and *PRON+VERB were prevalent among all the segments in the context-independent cases. Most of the patterns, excluding AUX+*VERB, that mainly occurred in context-independent TMs, were always very frequent among the context-dependent segments. This allows us to hypothesize that POS patterns may be useful indicators to identify potential context-dependent segments.

5.5. Summary

From the three experiments, we analyzed a total of 15,245 TMs for more than seven clients and from different domains. These were very frequently used TMs, one of the main criteria to select TMs used in the company. Overall, we were able to identify a total of 1,298 context-dependent TMs. From experiment 2 and experiment 3, we were able to distinguish two patterns that identified context-dependent segments, namely ADP+*PRON, for ES and ES-LATAM, and *VERB for PT and PT-BR.

We noticed that many of these context-dependent segments with gender constraints could be transformed into gender-neutral. Therefore, we proposed some suggestions on how to turn these context-dependent segments into gender neutral TMs (see chapter 7).

6. Error feedback loop and context-dependent TMs

A recurring task during the internship was the analysis of segments that were reported as context-dependent by the community of editors. The TMs were reported because they were blocked from edition, meaning that the editors could not perform the necessary changes to them. These segments totaled 89 for six different languages: German (DE), Italian (IT), Polish (PL), French (FR), Brazilian Portuguese (PT-BR) and Russian (RU), with English as the source language (see *Table 26*).

Language:	**Nº of segments:**
DE	63
IT	13
PL	5
FR	7
PT-BR	1
Total:	89

Table 26. *Reported segments*

In accordance with the data previously gathered, the reported segments presentent issues associated with gender and number agreement, register and also capitalization (*Table 27*). The latter category was not included in the ones proposed in the context annotation guidelines (see section 4.3). These cases occurred mainly in German, and some instances in Italian and Polish, and were related with clients' requests to have segments following a greeting in lowercase in an informal register. This segment usually corresponds to a sentence where the customer support agents introduce themselves. Even though this is not yet an established language standard, this category was validated by Unbabel's annotators, who were native speakers of these languages.

	Capitalization	Gender Agreement	Number agreement	Register	Total:
DE	61 (96%)	0	1 (2%)	1 (2%)	63 (70%)

IT	8 (62%)	5 (38%)	0	0	13 (14%)
PL	3 (60%)	0	1 (20%)	1 (20%)	5 (6%)
FR	0	7 (100%)	0	0	7 (9%)
PT-BR	0	0	0	1 (100%)	1 (1%)
Total:	72 (80%)	12 (14%)	2 (2%)	3 (3%)	89

Table 27. *Context-relates issues found*

All segments mentioned above were, indeed, context-dependent and, as a result, were unblocked from edition so that editors could perform the necessary changes to them. Consequently, and in conformity with the previous experiences, all of these segments with gender constraints were analyzed by the POS tagger, in order to test if the patterns found for the previous languages would be extended to these languages as well. The POS patterns found for the segments with gender constraints are shown in *Table 28*:

	AUX+*ADJ	VERB+*PRON	Total:
IT	3 (60%)	2 (40%)	5 (38%)
FR	7 (87%)	0	7 (62%)
Total:	10 (77%)	2 (15%)	12

Table 28. *POS patterns*

For IT data, three segments matched one of the patterns previously associated with context-dependent segments: AUX+*ADJ (97), and VERB+*PRON (98), respectively.

(97) EN: I would be very **grateful** if you would spend a minute filling the survey.
IT: Le [sarei **grato**] se potesse dedicare un minuto a compilare il sondaggio.

(98) EN: We may need to share this with the accommodation in order to confirm your claim.
IT: Potremmo aver bisogno di [condivider**lo**] con la struttura per confermare il suo reclamo.

In both examples, the adjective *grato* in (97) and the pronoun -*lo* in (98) are masculine, therefore, they are context-dependent.

As for FR data, the two patterns that matched with the segments were AUX+*ADJ (99):

(99) EN: I'll be glad to help you with that.
FR: Je [serai **heureux**] de vous aider avec cela.

Even though only a few segments were reported due to gender agreement errors for lack of context information, these results allow us to hypothesize that the POS patterns found could generalize for other languages that are morphologically marked by grammatical gender.

7. How to create gender neutral TMs

After a thorough analysis of extensie datasets with very diverse segments from different clients and different customer support domains, it was noticeable that gender agreement was the most prominent context related issue. As mentioned before, English is a language that has no gender assignment, therefore, is a gender neutral language. However, the same is not true for Romance Languages such as Portuguese or Spanish, where most words inflect in gender. Accordingly, having segments with gender information may compromise the meaning of the full text.

In this chapter, we will then present some suggestions on how to turn context-dependent segments with gender constraints into gender neutral TMs, thus allowing their use regardless of the context.

6.1. Portuguese and Brazilian Portuguese

For both languages, we found a total of 592 context-dependent TMs with gender constraints, 190 for PT and 402 for PT-BR. All of these segments can be divided into three main categories: participial form of verbs, pronouns and adjectives.

Firstly, the most common context related issue regarding gender was the use of participial forms of the verbs, such as "obrigado/obrigada" (thank you). *Table 1* shows the number of instances for this expression for all the datasets analyzed in this thesis.

	Experiment 1	Experiment 2	Experiment 3	Total:
PT	-	68	65	97
PT-BR	26	56	77	159
Total:	26	124	142	292

Table 1. *Distribution of the verb "Obrigado/obrigada"*

One proposal to solve this issue would be to replace this expression for a similar verb with the same meaning, however without gender constraints. This would be the case for the verb

Agradeço (“thank you”), which has an equal meaning but is gender neutral. Next are some examples (1 and 2):

(1) EN: **Thanks** for your patience.
    (a) Context-dependent: **Obrigado** pela sua paciência.
    (b) Gender neutral: **Agradeço** a sua paciência

(2) EN: Please accept my apologies for the delay in responding and **thank you** for your patience.
    (a) Context-dependent: Por favor, aceite minhas desculpas pelo atraso na resposta e **obrigado** pela sua paciência.
    (b) Gender neutral: Por favor, aceite minhas desculpas pelo atraso na resposta e **agradecemos** a sua paciência.

This participial verb occurred over 292 times in all datasets. Using the verb *Agradeço* instead would allow these curated TMs to be blocked from edition, thus reducing the amount of time editors spend on a task, and preventing the addition of non-necessary changes to an already high quality segment appropriate for a male or female addresser.

Pronouns were the second category with most occurrences for both PT and PT-BR. They were related to the pronouns *-lo* and *-la*. These pronouns are anaphoric in the sense that they establish an anaphoric relation with their antecedent. However, their antecedent does not occur in the same sentence, therefore, making these segments context-dependent. *Table 2* shows the distribution of these among the three experiments.

	Experiment 1	Experiment 2	Experiment 3	Total:
PT	-	28	15	43
PT-BR	4	14	70	88
Total:	4	42	85	131

Table 2. *Distribution of the pronoun -lo(a)*

As shown in the examples (3) and (4), our suggestion is to rephrase the segments without losing the original meaning but avoiding pronouns with gender constraints attached to them.

(3) EN: If you no longer have access to your phone number, please let us know and we'd be happy to provide you with further assistance on recovering your account.
    (a) Context-dependent: Se já não tem acesso ao seu número de telefone, por favor, avise-nos e teremos todo o prazer em ajudá-**lo** a recuperar a sua conta.
    (b) Gender neutral: Se já não tem acesso ao seu número de telefone, por favor, avise-nos e teremos todo o prazer em **prestar-lhe** mais assistência na recuperação da sua conta.

(4) EN: If there is anything else I can help you with, please do let me know.
    (a) Context-dependent: Se houver algo mais em que eu possa ajudá-**lo**, por favor me avise.
    (b) Gender neutral: Se houver algo mais em que eu possa ajudar, por favor avise-me.

In general, both PT and PT-BR make little use of adjectives. *Table 3* shows the adjectives distribution among the three experiments.

	Experiment 1	Experiment 2	Experiment 3	Total:
PT	-	2	13	15
PT-BR	6	4	76	86
Total:	6	6	89	101

Table 3. *Distribution of the adjectives*

After the analysis, we noticed that for PT, the common adjectives used were *interessado/a* (interested) and *descansado/a* (rested). For all of these cases, the adjectives can be replaced with an equivalent name, such as in (5b), or with a nominal expression such as in (6b).

(5) EN: If you're interested in re-subscribing you can do so at any time directly from the app or at URL-0.
    (a) Context-dependent: Se estiver **interessado** em voltar a subscrever, poderá fazê-lo a qualquer momento diretamente a partir da aplicação ou em URL-0.
    (b) Gender neutral: Se for do seu **interesse** voltar a subscrever, poderá fazê-lo a qualquer momento diretamente a partir da aplicação ou em URL-0.

(6) EN: Rest assured our team is working hard to fix this matter to avoid difficulties in the future when using our service.
    (a) Context-dependent: Fique **descansado** que a nossa equipa está a trabalhar arduamente para corrigir este assunto para evitar dificuldades no futuro ao usar o nosso serviço.
    (b) Gender neutral: **Tenha a certeza** de que a nossa equipa está a trabalhar arduamente para corrigir este assunto para evitar dificuldades no futuro ao usar o nosso serviço.

As for, PT-BR the most frequent adjectives used are *tranquilo/a* (calm), *preocupado/a* (worried) and *satisfeito*/*a* (satisfied). Similar to PT, these adjectives can be replaced with either an equivalent name, such as in (8b), or with a nominal expression, such as in (9b). Not all the adjectives found were cited, however the same solution may be applied to all.

(7) EN: Rest assured, there have been no discrepancies with the rewards.
    (a) Context-dependent: Fique **tranquilo**, não houve discrepâncias com as recompensas.
    (b) Gender neutral: **Tenha a certeza** de que não houve discrepâncias com as recompensas.

(8) EN: I understand that you are concerned about not receiving the rewards for inviting your friend.
    (a) Context-dependent: Eu entendo que você está **preocupado** em não receber as recompensas.

(b) Gender neutral: Eu entendo a sua **preocupação** em não receber as recompensas.

(9) EN: If you were satisfied with my assistance, please take the survey when you get one.
   (a) Context-dependent: Se você ficou **satisfeito** com a minha assistência, por favor, responda à pesquisa quando você receber uma.
   (b) Gender neutral: Se **apreciou** a minha assistência, por favor, responda à pesquisa quando você receber uma.

6.2. Spanish and Latin-america Spanish

For both language variants, a total of 671 segments with gender constraints were found, 331 for ES and 340 for ES-LATAM. For these languages, there were two problems: pronouns and adjectives. Unlike PT and PT-BR, these languages did not make use of verbs.

The common "problematic" pronoun was *nosotros* (us), which is a pronoun in its masculine form. *Table 4* shows the distribution of this pronoun through the experiments.

	Experiment 1	Experiment 2	Experiment 3	Total:
ES	14	45	122	181
ES-LATAM	-	30	61	91
Total:	14	75	183	272

Table 4. *Distribution of the pronoun "nosotros"*

All the TMs with this pronoun were similar to the following ones, therefore, this suggestion is applicable in all cases.

(10) EN: Thanks for contacting **us**.
   (a) Context-dependent: Gracias por contactar con **nosotros**.
   (b) Gender neutral: Gracias por **contactarnos**.

(11) EN: If there is anything else we could help you with, please do not hesitate to contact **us** again.

(a) Context-dependent: Si hay algo más en lo que te podamos ayudar, por favor, no dudes en ponerte en contacto con **nosotros** de nuevo.

(b) Gender neutral: Si hay algo más en lo que te podamos ayudar, por favor, no dudes en **contactarnos** de nuevo.

The remaining cases involving pronouns are related to the pronoun *-lo* (see *Table 5*) which is similar to PT and PT-BR.

	Experiment 1	Experiment 2	Experiment 3	Total:
ES	1	22	4	27
ES-LATAM	-	19	8	37
Total:	1	41	12	64

Table 5. *Distribution of the pronoun -lo*

(12) EN: I will surely try to help you with this issue.

(a) Context-dependent: Intentaré ayudar**lo** con este problema.

(b) Gender neutral: Intentaré ayudar**le** con este problema.

An adjective that generates contextual problems that was also very frequent among ES and ES-LATAM is *encantado* and *encantada*. *Table 3* shows the distribution of this adjective among the three experiments.

	Experiment 1	Experiment 2	Experiment 3	Total:
ES	30	7	39	76
ES-LATAM	-	39	104	143
Total:	30	46	143	219

Table 3. *Distribution of the adjective "encantado(a)"*

(13) EN: I'm happy to provide you with further information today.

(a) Context-dependent: Estoy **encantada** de proporcionarte más información hoy.

(b) Gender neutral: **Tengo todo el gusto** de proporcionarte más información hoy.

(14) EN: If you have any other questions, we are happy to help.

(a) Context-dependent: Si tienes cualquier otra pregunta, estaremos **encantados** de ayudarte.

(b) Gender neutral: Si tienes cualquier otra pregunta, **tendremos todo el gusto en ayudarte**.

Turning the category *Noun* into gender neutral is a somewhat arduous task. Since it was a gaming domain, all the nouns corresponded to the word *player* (*jogador* for portuguese and *jugador* for spanish). For these cases, the TMs could be unblocked in order to allow the editors to perform the necessary changes according to the context. However, one TM must be pointed out. The segment in (101) was found for PT. The translated text in Portuguese has the noun *senhor* or sir in English. However, the corresponding word in English is the pronoun *you*, making this segment context-dependent. A solution would be to use the second person pronoun *você*.

(100) Source: A checkmark will appear on your profile, signaling to potential matches that you're really **you.**

(a) Context-dependent: Vai aparecer uma marca de verificação no seu Perfil, a indicar a potenciais matches que realmente é **o senho**r.

(b) Gender Neutral: Vai aparecer uma marca de verificação no seu Perfil, a indicar a potenciais matches que realmente é **você**.

Creating gender neutral segments is an important step, as it would allow blocking more TMs, therefore reducing the time that annotators would spend post-editing these segments. In total, 15,245 TMs were analyzed in which 1,263 were context-dependent due to gender constraints. Our goal was not to eliminate every trace of gender in the segments, it was however

to have gender balanced segments by eliminating these types of constrictions in all frequent segments. Applying the suggestions mentioned above would reduce 8% of these problems and overall, increase the productivity in the post-editing process.

7. Conclusion and future work

The present thesis aimed at the creation of gender neutral and context-independent segments. In order to achieve this, a thorough analysis was conducted of three datasets consisting of Translation Memories from the customer support domain, analyzed regarding their (in)dependency of context. Our first goal was to understand what grammatical aspects were frequently involved in context-dependent TMs.

For the first experiment, we validated the classification that was previously done by the professional community, and analyzed a total of 2,045 TMs as isolated segments. From this, we obtained Register and Gender agreement as the most context-dependent categories. A direct result from this analysis was the reformulation of the first guidelines and development of a new version, more concise and unambiguous in order to provide clear instructions of the task. Our typology included the following categories: Gender agreement, Number agreement, Register, Ellipsis and Terminology. The guidelines were validated by the NLP analyst managing the linguistic documentation to support the communities and will be applied soon after the submission of this report.

We repeat the same procedures for the second experiment, however, using the new version of the typology and extending the dataset for 5,200 TMs, but this time taking into account the full context, as the TMs occurred in tickets (emails). In addition to the previous two categories, we found two new ones: Ellipsis and Terminology. However, gender was still the most context-dependent category, totaling 98% of all cases. Overall, for this first part of the analysis, we annotated a total of 338 context-dependent segments and the data was classified with a POS tagger. We were able to identify eight patterns: one exclusive for PT (*PRON+VERB), two exclusive for PT and PT-BR (*VERB and *VERB+ADP) and one for ES and ES-LATAM (ADP+*PRON) and the remaining were common for all language pairs (VERB+*ADJ, AUX+*ADJ, VERB+*PRON and AUX+*VERB).

For the third and final experiment, we wanted to test the patterns found, in order to verify if they were able to identify only context-dependent data. Therefore, once again we extended the dataset for a total of 8,000 segments that had never been classified and analyzed with the POS tagger.

One of our initial assumptions was that through POS information it would be possible to automatically identify all segments that were context-dependent. In the second analysis, we were able to isolate two very frequent and relevant patterns for context-dependent occurrences (*VERB and ADP+*PRON). Our experiments with POS taggers led us to conclude that the POS patterns are not always sufficiently discriminative between context-dependent and context-independent. Conducting a root-cause analysis, we notice that context-dependent segments involve specific words. For instance, the 3rd person singular pronouns for PT and PT-BR (*-lo* and *-la*) and 1st person plural pronouns for ES and ES-LATAM (*nosotros*) were very common and only knowing that these pronouns can occur is already very informative. We also encountered adjectives such as *satisfeito(a)*, *interessado(a)* or *encantado(a)* and *emocionado(a)*, and other similar adjectives that allow one to express appreciation or dissatisfaction and also specific participial verb forms such as *obrigado/obrigada*. The high frequency of this vocabulary on the customer support domain is so relevant that it may even overcome the POS tagger information. A future contribution would be to verify if very frequent lemmas may surpass the POS information or if both combined can add value to the improvements of a system detecting context (in)dependent TMs and generating gender neutral alternatives when possible, without ever compromising the meaning.

In total, we analyzed 15,245 TMs for more than seven clients and from different domains (i.e. gaming, technology). These were very frequently used TMs, one of our main criteria to select TMs. Overall, we were able to identify a total of 1,298 TMs in which 1,263 had gender constraints. We were able to suggest gender neutral TMs for the most frequent ones, therefore blocking for edition 8% of the data analyzed and maintaining the same meaning.

The work conducted in this thesis is being applied in the Error Feedback Loop at Unbabel and aligned with clients' reports and NLP modules. By blocking the most frequent TMs analyzed in our data, our contribution can be reflected in fewer errors in terms of gender.

Bibliography

ALPAC, 1966. Language and Computers in Translation and Linguistics. A Report by the Automatic Language Processing Advisory Committee.

Ariel, M. (2009). Discourse, grammar, discourse. Discourse studies, 11(1), 5-36.

Arthern, P. J. (1979). Machine translation and computerized terminology systems. Translating and the Computer. North-Holland.

Bawden, R., Sennrich, R., Birch, A., & Haddow, B. (2017). Evaluating discourse phenomena in neural machine translation. https://arxiv.org/pdf/1711.00513.pdf

Bowker, L. (2002). Computer-aided translation technology: A practical introduction. University of Ottawa Press. 92-127.

Bowker, L. and Fisher, D. (2010). Computer-aided translation. Handbook of translation studies, 1, 60-65.

Brown, G., & Yule, G. (1983). Discourse Analysis. Cambridge University Press.

Cai, X., & Xiong, D. (2020, December). A test suite for evaluating discourse phenomena in document-level neural machine translation. In Proceedings of the Second International Workshop of Discourse Processing (pp. 13-17).

Carnie, A. (2012). Syntax: A generative introduction. John Wiley & Sons. 43-66.

Carrozo, M. (2017). How Unbabel's "Translation as a Service" will translate everything to human quality. Available at: https://resources.unbabel.com/blog/translate-human-quality (Accessed: 21-12-2021)

Castilho, S., Camargo, J. L. C., Menezes, M., & Way, A. (2021, November). DELA Corpus-A Document-Level Corpus Annotated with Context-Related Issues. In *Proceedings of the Sixth Conference on Machine Translation*.

Chowdhury, G. (2003) Natural language processing. Annual Review of Information Science and Technology, 37. 51-89. https://strathprints.strath.ac.uk/2611/1/strathprints002611.pdf

Fernandes, P., Yin, K., Neubig, G., & Martins, A. F. (2021). Measuring and increasing context usage in context-aware machine translation. arXiv preprint arXiv:2105.03482.

De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2), 255-308.

Fishman, A. S. (1978). The effect of anaphoric references and noun phrase organizers on paragraph comprehension. Journal of Reading Behavior, 10(2), 159-170.

Forcada, M. L. (2017). Making sense of neural machine translation. Translation spaces, 6(2), 291-309.

Garcia, I. (2009). Beyond translation memory: Computers and the professional translator. *The Journal of Specialised Translation,* 12(12). 199-214.

Garcia, I. (2015). Computer-aided translation: systems. Routledge Encyclopedia of Translation Technology. 68-87.

Güngör, T. (2010). Part-of-Speech Tagging. *Handbook of natural language processing*, 2, 205-235.

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., & Loáiciga, S. (2018, October). A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings*

of the Third Conference on Machine Translation: Shared Task Papers. 570-577. https://aclanthology.org/W18-6435.pdf

Halliday, M. A. K. and Hasan, R. (1976). Cohesion in English (No. 9). Longman.

Halliday, M.A.K. (1985). Context of situation. In *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*. Deakin University. 3-12.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. https://arxiv.org/pdf/1803.05567.pdf?source=post_page----------

House, J. (2006). Text and context in translation. Journal of pragmatics, 38(3). 338-358.

Hutchins, W. J. (1995). Machine translation: A brief history. In *Concise history of the language sciences.* Pergamon.431-445.

Hutchins, W. J. (1998). The origins of the translator's workstation. Machine Translation, 13(4), 287-307.

Hutchins, W. J. (2001). Machine translation over fifty years. Histoire epistémologie langage, 23(1), 7-31. https://www.persee.fr/doc/hel_0750-8069_2001_num_23_1_2815

Jurafsky, D., and Martin, J. H. (2014). Speech and language processing. Vol. 3. US: Prentice Hall.

Kay, M. (1997). The proper place of men and machines in language translation. *machine translation*, 12(1), 3-23.

Kenny, D. (2018). Machine translation. In Routledge Encyclopedia of Translation Studies Routledge. 305-310.

Koehn, P. (2017). Neural machine translation. https://arxiv.org/pdf/1709.07809.pdf%2C2ndpublicdraft

Koehn, P. (2020). Neural machine translation. *Cambridge University Press*.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció, (12)*. 455-463.

Lopes, A. V., Farajian, M. A., Bawden, R., Zhang, M., & Martins, A. F. (2020). Document-level neural MT: A systematic comparison. In *22nd Annual Conference of the European Association for Machine Translation*. https://hal.archives-ouvertes.fr/hal-02900686/document

Malinowski, B. (1923) The problem of meaning in primitive languages. In *C. K. Ogden, & I. A. Richards (Eds.), The Meaning of Meaning. London: K. Paul, Trend, Trubner.* 296-336

Maruf, S., Saleh, F., & Haffari, G. (2019). A survey on Document-level Neural Machine Translation: Methods and Evaluation. https://arxiv.org/pdf/1912.08494.pdf

Melby, A. K., & Foster, C. (2010). Context in translation: Definition, access and teamwork. *Translation & Interpreting, The,* 2(2), 1-15. https://www.trans-int.org/index.php/transint/article/viewFile/87/70

Meyer, T., & Webber, B. (2013, August). Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*. https://aclanthology.org/W13-3303.pdf

Mitkov, R., Choi, S. K., & Sharp, R. (1995). Anaphora resolution in machine translation. In *Proceedings of the Sixth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*. https://aclanthology.org/1995.tmi-1.6.pdf

Müller, M., Rios, A., Voita, E. and Sennrich, R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

Reinke, U. (2018). State of the art in translation memory technology. *Language technologies for a multilingual Europe,* 4, 55. https://library.oapen.org/bitstream/handle/20.500.12657/28285/1001677.pdf?sequence=1#page=63

Sanchez, Marina (2020), If language is subjective, how can we measure translation quality? (Accessed: 21-12-2021) https://resources.unbabel.com/blog/measure-translation-quality

Shiwen, Y., & Xiaojing, B. (2014). Rule-based machine translation. In *The Routledge Encyclopedia of Translation Technology*. Routledge. 224-238.

Somers, H. (2003). Translation memory systems. *Benjamins Translation Library*, 35. 31-48.

Stein, D. (2018). Machine translation: Past, present and future. *Language technologies for a multilingual Europe, 4(5).*

Unbabel (2017). https://resources.unbabel.com/blog/scale-translation-quality-speed (Accessed: 24-12-2021)

Voita, E., Sennrich, R., & Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion.

Voita, E., Serdyukov, P., Sennrich, R., & Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. https://arxiv.org/pdf/1805.10163.pdf

Yin, K., Fernandes, P., Martins, A. F., & Neubig, G. (2021). When Does Translation Require Context? A Data-driven, Multilingual Exploration. https://arxiv.org/pdf/2109.07446.pdf

Yule, G., & Widdowson, H. G. (1996). Pragmatics. *Oxford university press*. 17-21.