universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Syntactic Diversity in Machine-Translated and Post-Edited Literary Texts

verfasst von | submitted by

Ramona Schultze BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 587
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Multilingual Technologies
Betreut von | Supervisor:	Assoz. Prof. Mag. Dr. Dagmar Gromann BSc

Acknowledgements

Thank you to my supervisor, Assoz. Prof. Mag. Dr. Dagmar Gromann, BSc, for her support and constructive feedback throughout. Another Thank You to Ass.-Prof. Mag. Dr. Waltraud Kolb and Univ. Ass. Gábor Recski, PhD, who brought this topic to my attention and made the data available. I owe my deepest gratitude to my family, my friends, and especially my library buddies, you know who you are, as well as my colleagues from the MLT program. The past two years have been an unforgettable experience, made all the more enjoyable and manageable thanks to you.

Abstract

Machine-generated and post-edited translations differ from their equivalent counterpart, translations done by humans. These differences reveal themselves, among others, upon examining their syntactical structure, suggesting a more diverse product for the latter. Previous works on syntactic equivalence across the different types of translation, namely Human Translation (HT), Post-Edited Machine Translation (PEMT), and Machine Translation (MT), consist of a variety of source text genres as objects of the investigation. Although literary texts have been part of some, this thesis aims to look at the translations of literary texts exclusively to uncover systematic differences between the grammatical dependencies present in each type of translation. Finding these patterns or consistent variations in the grammatical structures and syntactic characteristics reveals the strengths and weaknesses of each translation method and helps in understanding how the translation method affects the syntactic complexity. State-of-the-art syntactic parsers from the Universal Dependency (UD) framework are used to process the translations of a Brazilian Portuguese-to-German literary dataset of 24 short stories by Lima Barreto and Machado de Assis to gain insights into their quantitative and qualitative properties. The texts were preprocessed, sentence- and word-aligned, and evaluated using multiple measures of syntactic diversity, complemented by qualitative inspection. The results show that MT remains structurally close to the Source Text (ST), while HT involves greater syntactic reorganization. The PEMT does not simply fall between these two modes but instead exhibits a distinct syntactic profile, more varied than MT, yet not fully convergent with the strategies of HT.

Zusammenfassung

Maschinell erzeugte und post-editierte Übersetzungen unterscheiden sich von Übersetzungen, die von Menschen angefertigt wurden. Diese Unterschiede zeigen sich unter anderem in der syntaktischen Struktur, wobei sich für die menschliche Übersetzungen eine größere Vielfalt beobachten lässt. Frühere Arbeiten zur syntaktischen Äquivalenz zwischen den verschiedenen Übersetzungen, also menschlich angefertigt, maschinell übersetzt und Post-Editierte Maschinelle Übersetzung (PEMÜ), untersuchen eine Vielzahl unterschiedlicher Textgenres als Ausgangsmaterial. Obwohl literarische Texte teilweise berücksichtigt werden, konzentriert sich die vorliegende Arbeit ausschließlich auf literarische Übersetzungen, um systematische Unterschiede in den grammatischen Abhängigkeiten der jeweiligen Übersetzungsarten aufzudecken. Die Identifikation solcher Muster oder konsistenter Variationen in den grammatischen Strukturen und syntaktischen Eigenschaften zeigt die Stärken und Schwächen der einzelnen Übersetzungsmethoden auf und trägt zum Verständnis bei, wie die Wahl der Methode die syntaktische Komplexität und Diversität beeinflusst. Mithilfe moderner Parser des Universal Dependency (UD) Frameworks werden Übersetzungen eines portugiesisch-deutschsprachigen literarischen Korpus von 24 Kurzgeschichten von Lima Barreto und Machado de Assis untersucht, um Einblicke in deren quantitative und qualitative Eigenschaften zu gewinnen. Die Texte wurden für die Verarbeitung vorbereitet und satz- und wortweise ausgerichtet, um sie dann mithilfe mehrerer Maße zur syntaktischen Diversität auszuwerten. Qualitative Analysen ergänzen dies. Die Ergebnisse zeigen, dass maschinelle Übersetzungen strukturell eng am Ausgangstext bleiben, während Menschliche Übersetzungen (MÜ) eine stärkere syntaktische Umstrukturierung aufweisen. Die PEMÜ nimmt dabei keine Mittelstellung zwischen beiden ein, sondern zeigt ein eigenes syntaktisches Profil, vielfältiger als die MÜ, jedoch ohne die Strategien der menschlichen Übersetzung vollständig zu übernehmen.

Contents

List of Tables

List of Figures

1. Introduction

Studies have shown that Machine Translation (MT) output tends to exhibit lower syntactic and lexical diversity compared to Human Translation (HT) (Vanmassenhove et al., 2021; Baldo de Brébisson, 2019; Recski and Kádár, 2023). Similarly, Post-Edited Machine Translation (PEMT) often retains syntactic structures that are closer to the source language, suggesting less divergence in sentence structure (Toral, 2019). These features and patterns, among others, that are found in MTs and PEMTs but not in HTs are known as *machine translationese* and *post-editese* (Daems et al., 2017; Vanmassenhove et al., 2021). Both are derived from *translationese*, which expresses the differences found in translated texts compared to the original texts (Gellerstam, 1986).

Previous studies have analyzed the mentioned phenomena in different language pairs, on a diverse set of data, concerning genres or length, and focused on numerous aspects. Recski and Kádár (2023) measure the language complexity of translated texts using statistical measures such as the Type-to-Token Ratio (TTR), Herdan's C, and the Measure of Textual Lexical Diversity (MTLD). These metrics quantitatively evaluate lexical richness and syntactic complexity. Additionally, readability metrics, such as Gunning's Fog index, provide insight into text accessibility by analyzing sentence length and the proportion of complex words.

Another statistical analysis included the Part-Of-Speech (POS) sequences, a linguistic feature revealing that the output of PEMT aligns more closely with the source language (Toral, 2019). This suggests a stronger influence from the Source Text (ST) in PEMT than a more natural composition of the POS sequences in an HT. In contrast, qualitative approaches, such as error analysis, focus on identifying and categorizing errors in translations. These include grammar, style, adding or omitting information, localization issues, and morphosyntactic errors, as shown by Baldo de Brébisson (2019).

While these observations have been explored in various translation domains, this thesis aims to investigate syntactic diversity exclusively on a set of literary texts by using state-of-the-art syntactic parsing. A quantitative and qualitative analysis will be conducted to reveal systematic differences between the grammatical dependencies in the different translations. The starting point is a dataset of Brazilian Portuguese-to-German short stories that include three modalities: MT, PEMT, and HT. This thesis will focus on syntactic differences across these types of translations, leaving out lexical choices.

In order to make this investigation quantifiable across many languages, a unified and consistent framework is needed. Universal Dependencies (UD) is a still evolving but robust framework that offers a cross-linguistically consistent morphosyntactic annotation for 104 languages, also seeking to enable research in the Natural Language Processing (NLP) context (Zeman et al., 2024). By employing universal categories and language-independent guidelines, UD enables comparative syntactic analysis across languages through a simplified yet comprehensive representation of sentence structure and dependencies. Using UD models for syntactic parsing, this study will generate automatic annotations that reveal the dependencies and syntactic structures in the dataset. These annotations will form the basis for assessing syntactic diversity across the three translation types. The findings aim to contribute to a deeper understanding of how MT and PEMT handle the complexity of literary texts, offering insights into their suitability for creative and linguistically rich content, particularly in light of a potential lack of syntactic diversity. The data used for this analysis cannot be shared due to copyright reasons but the code is available at this GitHub repository[1].

1.1. Research Questions and Assumptions

The central research question of this thesis is *Which similarities and differences can be observed in terms of syntactic diversity in MT, PEMT and HT of literary texts translated from Brazilian Portuguese into German?*. In addition, a related sub-question is whether systematic patterns emerge in the grammatical dependencies of the different translation types and whether these patterns can be captured through quantitative, syntax-based evaluation metrics. By addressing these questions, this thesis aims to contribute to a clearer understanding of how the transitional modality affects the syntactic profile of literary texts. Based on previous research (Toral, 2019; Vanmassenhove et al., 2021), several assumptions guide the analysis. First, MT and PEMT outputs are expected to remain syntactically closer to the ST than the HT output, reflecting the tendency of MT to reproduce source structures and of PEMT to preserve many of these patterns despite human intervention. Although Post-Editing (PE) introduces human modification, the degree of syntactic restructuring is assumed to be lower than that in HTs produced from scratch. Second, HT is expected to display greater syntactic diversity, for example, through a longer and more complex sentence structure, reflecting translators' stylistic choices and the creative demands of literary texts. Finally, it is assumed that PEMT will show an intermediate profile, positioned between MT and HT, but not identical to either of them. These assumptions provide a starting point for the empirical investigation,

[1] `https://github.com/schultzer20/syntactic-diversity-thesis`

while the analysis aims to test and potentially refine current understandings of syntactic diversity in translation.

1.2. Motivation and Objectives

This thesis is motivated by the need to better understand the role of syntactic diversity in literary translation. Literary texts often rely on stylistic and structural variation to create meaning, rhythm, and aesthetic effect (Boase-Beier, 2011). A lack of syntactic diversity, as observed in MT output, therefore raises question about the adequacy of machine-generated or post-edited translations in this domain. While previous studies (Shaitarova et al., 2023; Volkart and Bouillon, 2024) have explored syntactic properties of MT, HT, and PEMT in various domains, literary texts remain underexplored. Addressing the gap is essential in order to evaluate whether MT and PEMT are capable of reproducing the syntactic richness characteristic of human literary translation. The objective of this study is therefore to provide a detailed analysis of syntactic diversity across three translation modalities, combining quantitative metrics with qualitative inspection. By applying automated measures such as dependency label changes and crossing alignments, the study seeks to quantify the extent to which each modality restructures syntax. These quantitative results will be complemented by a manual examination of selected examples, ensuring that the findings are not only statistically robust but also meaningful in the context of literary translation. In this way, the thesis aims to shed light on the syntactic behavior of MT and PEMT in comparison to HT, to assess their suitability for creative translation contexts, and to contribute to broader methodological discussions on how syntactic diversity can be measured in translation studies. Such an analysis and investigation is interesting for literary translators as well as translators of other creative text types. Furthermore, the findings might be interesting for developers of machine translation systems, providing insights into one area of potential and potentially substantial improvement of current approaches and systems. The findings might also be interested for customers commissioning post-editing tasks of creative text types by showing the potential syntactic impact in contrast to human translation. Finally, the results might be interesting within an educational context by showing clearly and providing use cases of syntactic differences between the different modes of translation.

2. Translation of Literary Texts

Literary texts are a distinct category of written expression that encompasses various genres, including novels, short stories, poetry, plays, and graphic literature (Hadley et al., 2022). Unlike technical or scientific texts, which primarily serve to convey precise information, literary texts emphasize artistic expression, emotional resonance, and cultural reflection which means that their primary function is aesthetic rather than purely informational (Hadley et al., 2022). Furthermore, literary texts have increasingly been recognized as "embody[ing] a state of mind" (Boase-Beier, 2011, 46). While non-literary texts may also exhibit individual style and reflect specific perspectives, these elements are far more pronounced in literary works according to Wright (2016). Thus, literary translation is widely recognized as a distinct and specialized domain within Translation Studies (TS), which encompasses the translation of genres such as fiction, poetry, drama, and literary non-fiction. It is fundamentally different from the translation of technical, scientific, or informational texts where clarity, terminological precision, and standardization dominate (Borg, 2022; Hadley et al., 2022). A literary translator has to make countless deliberate decisions about tone, rhythm, register, syntactic form, and stylistic nuance which make the process a highly creative and interpretive act and can be seen as a hybrid of close reading and writing (Hadley et al., 2022; Wright, 2016; Nelson and Maher, 2013). These choices are not incidental but central to preserving what is often referred to as the literariness or style of a text.

This chapter begins by examining the defining characteristics of literary texts and the specific challenges they pose for translation, in order to clarify the object of investigation for this thesis. It then presents key conceptual frameworks and strategies developed in literary translation studies to address these challenges and to conceptualize translation as a situated, interpretive, and culturally embedded practice. In addition, the chapter outlines relevant computational approaches, introducing core NLP concepts that provide the methodological basis for the quantitative analysis of translated literary texts in this study.

2.1. Characteristics and Challenges

Literary texts exhibit a range of characteristics and elements that distinguish them from other types of texts in the context of translation (Schogt, 1988; Wright, 2016).

These features significantly influence the translator's approach and the strategies they must employ to convey both meaning and artistic intent effectively (Wright, 2016).

One defining aspect of literary texts is their high degree of stylistic complexity (Wright, 2016). Style is a crucial element in literary works, as it determines how something is expressed and why a text is structured in a specific way (Schogt, 1988; Youdale, 2020). This includes the use of rhetorical devices such as metaphors, irony, alliteration, and allusions, but also choices about register or tone (Nelson and Maher, 2013). The translator must be able to recognize, interpret, and adequately reproduce these stylistic features in the target language, often requiring creative adaptation, as literary translation involves not just transferring meaning but also maintaining the artistic choices and stylistic expressions made by the author (Youdale, 2020; Wright, 2016).

Another essential characteristic is the inherent ambiguity and interpretative openness of literary texts (Wright, 2016). As Wright (2016) and Borg (2022) state, literary texts frequently contain implicit meanings that require the reader to actively engage in interpretation. Unlike technical texts, where clarity and precision are paramount, literary works often rely on suggestion and nuance. The translator must preserve this ambiguity rather than reduce the text to a single interpretation, ensuring that multiple readings remain possible in the translated version. This requires not only linguistic skill but also an awareness of literary analysis and the broader cultural or philosophical implications and discourse embedded in the text (Nelson and Maher, 2013).

Cultural boundaries also play a fundamental role in literary texts. The STs can be deeply rooted in the cultural context of their origin, often incorporating references to historical events, social norms and customs, and literary traditions that might be unfamiliar to readers in the target culture (Schogt, 1988). The translator must find ways to convey these cultural references, either by providing explicit explanations or by using culturally equivalent expressions in the target language, maintaining faithfulness to the ST and culture while ensuring accessibility to the target audience (Wright, 2016). This requires extensive knowledge of both the source and the target cultures, as well as a strategic approach in determining when and to what extent to domesticate the translation to maintain the intended impact of the ST (Nelson and Maher, 2013).

Another element to consider is the emotional factor. Literary texts aim to evoke emotions and create an aesthetic experience for the reader (Wright, 2016). As Schogt (1988) notes, particularly in poetry, the phonetic and rhythmic qualities of a language are integral to its meaning and impact. Maintaining this emotional and artistic effect in the target language often requires innovative solutions and a keen sense of linguistic rhythm and tone. The translator must balance accuracy

with the need to replicate poetic elements, such as meter, rhyme, and musicality, sometimes requiring significant deviations from the literal meaning to preserve the artistic integrity of the work in the target language (Nelson and Maher, 2013).

Closely related to the previous feature is the creativity in language use in literary texts. Authors may break grammatical rules and stylistic conventions to develop unique expressions and innovative linguistic structures (Youdale, 2020; Hadley et al., 2022). Embracing these creative liberties and finding equivalent unconventional expressions in the target language may involve deviating from strict literal translation to retain the artistic and expressive qualities of the ST (Borg, 2022). This need for creativity also applies to handling the absence of direct equivalences between STs and Target Texts (TTs). Depending on the used languages, direct one-to-one correspondences sometimes do not exist, forcing the translator to make creative decisions that best convey the spirit of the original(Wright, 2016). This aspect of translation emphasizes the need for flexibility and innovation, as strict adherence to source structures may diminish the artistic effect of the translated work.

Furthermore, literary translation is inherently subjective, as the process is shaped by the translator's individual interpretation, background, and stylistic preferences (Wright, 2016). Each translator brings a unique perspective to the text, leading to different possible translations of the same work. This subjectivity underscores the idea that there is no single correct translation, but rather multiple valid renditions that capture different aspects of the original (Wright, 2016). This subjectivity can be influenced by the translator's cultural background, their familiarity with the author's style, and their understanding of literary traditions, making literary translation an intellectual as well as a linguistic endeavor. Finally, literary translation can be seen as a creative negotiation between the ST and the target language while incorporating the translator's voice (Nelson and Maher, 2013). The translated work becomes a reinterpretation that carries traces of the original, but also reflects the creative input of the translator. The translator must act as both a linguist and an artist, making decisions that balance fidelity to the source with the aesthetic expectations of the target readership.

These characteristics and their arising challenges make literary translation a demanding and highly intricate task that requires extensive linguistic, cultural, and creative competence. The translator must engage deeply with the text and critically reflect on their choices to produce a translation that captures not only the meaning but also the artistry of the original work. This dynamic interplay between linguistic precision and creative adaptation highlights the complex nature of literary translation as both a scholarly and artistic endeavor.

2.2. Translationese and Translation Universals

As concepts related to the phenomenon of translationese will be examined in more detail in subsequent chapters, this section offers a brief overview of the term and its theoretical underpinnings as a foundation for the discussion. Understanding the nature of translationese is essential for analyzing the linguistic and stylistic characteristics of translated texts, especially in light of this thesis's focus on syntactic variation across human and machine translation outputs of literary works.

The term translationese was originally coined by Gellerstam (1986) to describe the distinct linguistic traces that frequently occur in translated texts and set them apart from texts originally composed in the same language. These characteristics are not merely the result of interference from a specific source language, but rather reflect deeper constraints inherent to the process of translation itself. Building on this observation, the concept of translation universals was first proposed by Baker (1993) to describe linguistic tendencies that appear consistently across translated texts, regardless of the language pair, genre, or translation direction.

Baker (1993) states that among the most widely discussed translation universals is a form of explicitation, the tendency to make implicit meanings more explicit in the target text by adding cohesive devices, clarifying referents, or providing background information. This is closely related to disambiguation and simplification, where ambiguous or structurally complex elements in the ST are rendered more transparently or reduced in syntactic complexity in translation. Another frequently observed tendency is normalization or conventionalization, whereby translated texts conform more strictly to the grammatical, lexical, or stylistic norms of the target language than texts written in the first language. Other recurring traits include the avoidance of repetition, the overuse of cohesive markers, and a greater display of target language conventions. In some cases, these tendencies result in what has been described as a third code (Frawley, 1984, 168), a hybrid linguistic register that is neither fully typical of the first language nor entirely foreign, but instead shaped by the unique pressures of translation.

Importantly, Baker (1993) emphasizes that these features should not to be understood as errors or deficiencies, but as manifestations of the specific constraints and goals of translation as a communicative act. Nonetheless, the presence of translationese can influence reader perception and stylistic quality, especially in genres such as literature, where the expectations for voice, creativity, and naturalness are particularly high (Youdale, 2020). For this reason, the study of translationese and translation universals, especially through corpus-based and syntactic analyses, remains central to evaluating both human and machine-generated translations and understanding the linguistic footprint that translation leaves behind (de Clercq et al., 2021).

2.3. Strategies

Literary translation does not follow a unified set of methods or fixed strategies, but instead relies on a wide range of approaches that depend on the ST and the target group and are shaped by personal interpretation and creativity (Schogt, 1988). These strategies are not strict formulas but flexible tools to navigate the unique challenges posed by literary texts. The notion of equivalence, be it formal or dynamic, plays a central role. The former focusing on preserving surface structure, whereas the latter seeks to evoke a similar response in the target readers as the ST did in its respective recipients (Wright, 2016). This equivalence can manifest itself across multiple levels such as lexical or grammatical.

One recurring theme is the translator's negotiation between literal and free translation, which, in other words, is a decision between translating word for word or more sense for sense (Washbourne and Van Wyke, 2019). The first emphasizes fidelity to content and form of the ST, even down to punctuation. In contrast, the latter prioritizes stylistic and functional resonance in the target language.

As Venuti (1995) brought forward, domestication and foreignization are opposite ends of a spectrum of strategies and methods applied during the translation process. A domesticated translation adapts the ST to the norms, expectations, and stylistic conventions of the target culture, often smoothing over cultural differences to make the text appear as if originally composed in the target language. In contrast, a foreignized translation preserves elements of the source culture and language, intentionally making the recipient aware of the foreign origin of the text, even at the expense of fluency or familiarity. This can include the presence of further explanations, glossaries, or footnotes (Youdale, 2020). These two strategies represent poles on a continuum of translation approaches and carry significant ethical and ideological implications (Venuti, 2018; Youdale, 2020). Although domestication may facilitate accessibility and commercial viability, foreignization challenges dominant cultural norms by exposing readers to otherness and cultural difference. In practice, however, most literary translations combine elements of both strategies, and the translator's choices are often shaped by their own ideological stance, the publishing context, and the intended readership (Wright, 2016).

On another more creative note, translators may generate and compare Alternative Translation Solutions (ATSs) that can include non-literal but more idiomatic options next to literal options and a collection of synonyms to chose from (Borg, 2022). Some ATSs may explain ambiguity while others keep it. The selection among these variants may be guided by auditory perception, stylistic considerations, or the perceived cultural function of the passage (Borg, 2022).

In addition to structural and stylistic considerations, literary translators must also approach the text through pragmatic and cognitive lenses. This involves analyzing not only the linguistic form but also the implicit meanings, authorial intent, and

contextual nuances embedded in the ST (Washbourne and Van Wyke, 2019). Translators engage in inferential processes and must negotiate the world knowledge of the source culture in relation to the target audience's cultural and cognitive environment (Hadley et al., 2022). As a result, strategies such as explicitation or adaptation may be employed to render implicit content more accessible or to recreate contextually situated sequences in a way that resonates with target readers. Beyond these broad tendencies, translators employ a wide variety of specific procedures. For instance, in handling Culture-Specific Items (CSIs), scholars have proposed numerous taxonomies that distinguish between source-language-oriented strategies, e.g., adaptation of the orthographie used in the ST or repetition, and target-language-oriented ones, e.g., deletion or synonymity (Youdale, 2020). Whether a translator chooses to conserve, adapt or substitute such elements depends on a number of factors such as their narrative function or their stylistic role, both individually and collectively.

In sum, literary translation involves a complex interplay of strategies, many of which defy strict categorization. Translators operate within shifting constraints, balancing fidelity and creativity, visibility and transparency, cultural representation and literary effect. Rather than following fixed rules, they engage in a situated, interpretive process shaped by text, context, and their own translational poetics. Within this dynamic, domestication and foreignization represent overarching orientations rather than rigid methods. Notably, both approaches may coexist within a single translation, reflecting the translator's shifting strategies across different textual moments.

2.4. NLP Basics

In order to analyze natural language computationally, raw text must first undergo a series of preprocessing steps that convert unstructured language data into structured representations interpretable by Machine Learning (ML) models. These foundational steps, such as tokenization, lemmatization, Part-Of-Speech (POS) tagging, Named Entity Recognition (NER), and syntactic parsing, not only enable basic linguistic analysis, but also serve as essential building blocks for more complex downstream tasks such as MT, semantic analysis, or translation quality evaluation. This subchapter outlines the core components of NLP, emphasizing their function, methodology, and relevance to cross-linguistic and translation-focused NLP applications.

Tokenization is the starting point for preprocessing a text for many tasks in the field of NLP (Park et al., 2020; Song et al., 2021; Webster and Kit, 1992). It is the process of segmenting a text into smaller units which carry contextual information and enable subsequent linguistic or computational analysis. Without a clear and

consistent representation of such units, any NLP operation is "impossible" (Webster and Kit, 1992, 1106). The goal of tokenization is to segment longer sequences of text into individual tokens, which can be words, or more commonly in neural approaches, subword units. The most basic form of a token is a word. Identifying words appears relatively straightforward in languages such as English, where whitespaces are used as explicit delimiters between words, but it can pose significant challenges in different languages. Concerning languages without explicit delimiters, such as Chinese or Korean, sophisticated algorithms capable of recognizing collocation patterns and lexical boundaries are required (Webster and Kit, 1992). Furthermore, the question arises how to handle more complex tokens such as idioms, fixed expressions, or Multi-Word Tokens (MWT). Depending on the NLP task, it may be important to split or preserve these expressions as indivisible tokens, e.g., in the case of MT a decomposition of an idiom can lead to errors in the Target Text (TT). Various tokenization strategies have been developed to address linguistic complexity across languages. These range from simple whitespace-based methods to more advanced approaches such as subword tokenization, and techniques that integrate morphological, collocational, or syntactic information to improve segmentation accuracy (Song et al., 2021; Schmidt et al., 2024; Webster and Kit, 1992; Park et al., 2020). In summary, tokenization is a crucial step in NLP that prepares text for computational processing.

Lemmatization is the linguistic process of normalizing words by reducing them to their base form, also called lemma, typically as it would appear in a dictionary. The lemma serves as the canonical form of a word, such as the infinitive for verbs or the singular nominative form for nouns. As a key step in linguistic preprocessing, lemmatization aims to reduce the variability of word forms across inflections, simplifying tasks such as Information Retrieval (IR), knowledge extraction, or semantic analysis (Eger et al., 2016). The process can be approached through different techniques, including rule-based methods using lexicons, pattern-based transformations, or neural models that generate lemmas character-by-character conditioned on context and POS information (Dorkin and Sirts, 2023; Eger et al., 2016). As Eger et al. (2016) states, rule-based systems are reliable for known word forms, but often struggle with ambiguity and irregularities, especially in morphologically rich languages, such as German or Latin. Context-sensitive models, on the other hand, are better equipped to handle rare forms and disambiguation. However, lemmatization is not without challenges. These include inconsistent lemma conventions, irregular forms, and limitations in handling historical variants or typographic errors (Eger et al., 2016; Dorkin and Sirts, 2023). Nonetheless, lemmatization contributes significantly to reducing data sparsity, enhancing search systems, and improving downstream syntactic and semantic tasks.

Stemming is a heuristic process that strips words down to their root forms by

removing derivational and inflectional suffixes. Unlike lemmatization, stemming does not always yield a valid dictionary word, and instead aims to group morphologically related words under a shared stem. Stemming algorithms range from iterative suffix removal to longest-match and context-sensitive variants, some of which involve multi-phase processing to address spelling variants (Lovins, 1968). While efficient and computationally less demanding, stemming can lead to errors due to overgeneralization or ambiguity (Slawik et al., 2015). For instance, suffixes may be mistakenly removed from stems, or two unrelated words might be conflated due to superficial similarity. Despite its limitations, stemming remains useful for tasks where high linguistic precision is not required and the reduction of word form variability is sufficient. It helps reduce index size in search systems, increase retrieval recall, and can serve as a lightweight preprocessing step in a range of NLP applications (Slawik et al., 2015; Lovins, 1968).

POS Tagging is the process of assigning each token in a sentence its grammatical category, such as preposition, noun, verb, or adjective. This syntactic annotation step is foundational in NLP as it provides structural insight into a sentence and supports downstream tasks such as parsing, lemmatization, and Information Extraction (IE) (Straka et al., 2016; Sahala and Lindén, 2023). While basic POS-tags indicate the word class, extended tagsets in morphologically rich languages also encode grammatical features, such as case, gender, number, and tense (Eger et al., 2016). POS-tagging methods range from early rule-based and statistical models to modern neural architectures. Contemporary systems, such as Stanza (Qi et al., 2020), often use ensembles that combine lexical and neural features. Despite advances, challenges remain in handling lexical ambiguity, Out-Of-Vocabulary (OOV) tokens, and inconsistent annotations (Sahala and Lindén, 2023; Tseng et al., 2005; Banko and Moore, 2004). Nevertheless, POS-tags are central to many NLP applications, including parsing, search, disambiguation, language modeling, NER, and MT.

Named Entity Recognition is a core NLP task that focuses on detecting and categorizing named entities within a text, including organizations, locations, people, dates, and other specific references (Manning et al., 2014). The goal is to extract structured data from unstructured text, enabling applications such as IR, Question Answering (QA), and improving MT. NER systems rely on rule-based, statistical, or neural methods. Hybrid approaches often combine pattern recognition and neural inference for increased robustness (Jiang et al., 2016; Lample et al., 2016). Several challenges persist in NER, such as entity ambiguity, domain-specific variations, or even obtaining high-quality training data (Straková et al., 2014; Jiang et al., 2016; Satoshi Sekine, 2009). Additionally, recognizing rare or OOV entities, especially in morphologically complex languages, and multiple occurrences in various forms remains difficult (Satoshi Sekine, 2009). Context plays a vital role in resolving

these ambiguities. To conclude, NER aids in identifying relevant entities in QA systems, enriches knowledge bases, and improves sentiment analysis by attributing opinions to entities. Moreover, NER can enhance MT by ensuring accurate and consistent rendering of named entities across languages.

Dependency Parsing is a core method in syntactic analysis that represents the structure of a sentence as a collection of binary, directed grammatical relations, also called dependencies, between individual words. Each word is linked to a syntactic head, which can be another word in the sentence or a special root symbol, forming a dependency tree in which each arc is labeled with a grammatical function, such as subject, object, or modifier as depicted in Fig. 1 (Jurafsky and Martin, 2025). This format is particularly effective in modeling predicate-argument structure and supports tasks such as QA, IE, and coreference resolution (Jurafsky and Martin, 2025). In a nutshell, dependency parsing models binary relations between individual words and captures head-dependent relations directly.

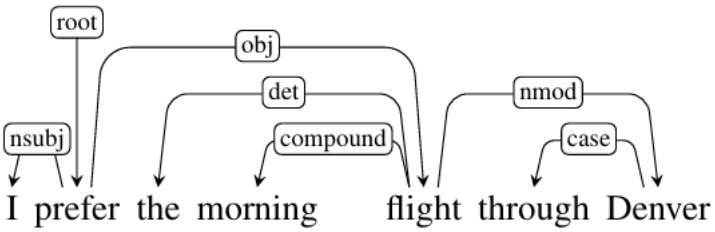

Figure 1.: Dependency parsing formalism by example sentence (Jurafsky and Martin, 2025, 411)

Constituency Parsing or phrase structure parsing, analyzes the hierarchical organization of phrases in a sentence by recursively grouping words into nested syntactic units such as Noun Phrases (NP) or Verb Phrases (VP), which capture the complete phrase structure and syntactic categories of a sentence (Jurafsky and Martin, 2025). Thus, constituency parsing emphasizes structural grouping and focuses on constituent boundaries and syntactic categories. It is grounded in formal grammars, such as Context-Free Grammar (CFG) and Head-Driven Phrase Structure Grammar (HPSG), the latter of which allows the integration of constituency and dependency representations (Zhou and Zhao, 2019). A CFG is a formal system that is used to model the constituent structure of natural language (Jurafsky and Martin, 2025). It defines how sentences can be generated and structurally analyzed by specifying rules for combining words into phrases and sentences. A CFG consists of a set of non-terminal symbols, e.g., NP, terminal symbols meaning the actual words, production rules that describe how symbols can be expanded, and a start symbol from which sentence generation begins. CFGs are used to generate valid sentences in a language and to assign hierarchical structure

to sentences through syntactic parsing. The structure is expressed in parse trees. These trees are typically annotated in treebanks, which serve as training data for syntactic parsers. While CFGs offer a powerful way to represent sentence structure, they also face limitations such as structural ambiguity, where a sentence can have multiple valid parse trees, especially with Prepositional Phrase (PP) attachment constructions as illustrated in Fig. 2 (Jurafsky and Martin, 2025).

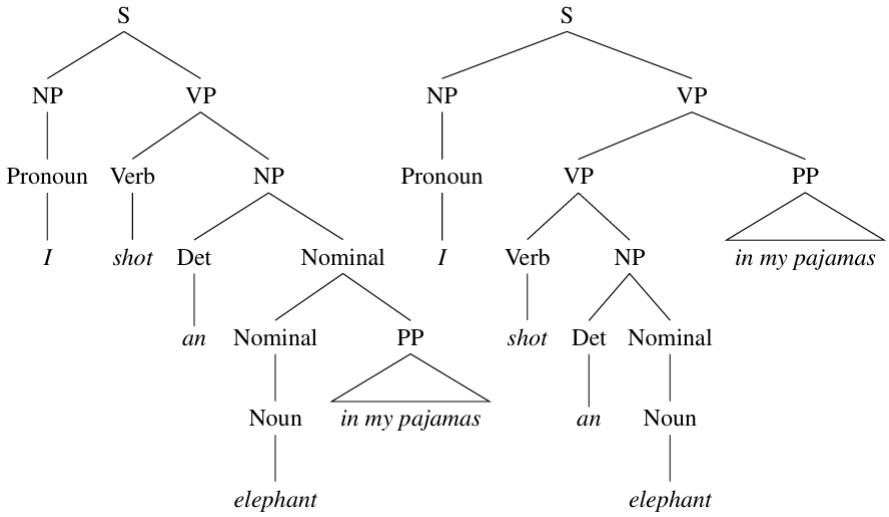

Figure 2.: Two possible constituency parse trees of the same sentence. Left: the elephant is in pajamas. Right: the subject 'I' is in pajamas (Jurafsky and Martin, 2025, 369)

Representation of Annotation. Linguistic annotations provide the structural foundation for a wide range of NLP tasks, including syntactic parsing, NER, and semantic role labeling (Jurafsky and Martin, 2025). These annotations encode linguistic features, such as token identity, grammatical relations, or semantic categories, in structured formats that facilitate automatic processing and evaluation. A widely adopted cross-linguistic framework for such annotations is UD which standardizes syntactic and morphological labels to ensure interoperability across languages and tools (Üstün et al., 2020). UD annotations are typically stored in the CoNLL-U format, a ten-column, tab-separated structure that includes information on token form, lemma, the universal and the language-specific POS-tag, morphological features, syntactic head index, and dependency relation (Straka et al., 2016). These annotations serve as training data for parsing models and as a consistent basis for evaluating syntactic shifts in MT or other multilingual tasks. In contrast, constituency parsing represents sentences as hierarchical phrase

structure trees, commonly following the bracketed format of the Penn Treebank (PTB) (Jurafsky and Martin, 2025). These trees reflect the internal structure of syntactic constituents, e.g., NP, VP, offering a better fit for capturing syntactic alternations and phrase-level transformations (Zhou et al., 2020). The HPSG framework provides a unified representation that integrates both syntactic and semantic information, enabling joint modeling of constituency and dependency structures which further illustrate the potential of combining both paradigms for improved syntactic analysis and cross-linguistic applicability (Zhou et al., 2020; Zhou and Zhao, 2019).

3. Machine Translation Preliminaries

This chapter provides the foundational concepts necessary to understand the technical and theoretical underpinnings of MT. It begins by outlining the main types of MT systems, including rule-based, statistical, and neural approaches. Subsequently, the chapter introduces key principles from ML that underpin modern MT models, such as supervised, unsupervised, and reinforcement learning. A central focus is placed on neural architectures, including feed-forward and recurrent networks, encoder-decoder structures, attention mechanisms, and the Transformer architecture. Additionally, the chapter explores the emergence of Large Language Models (LLMs) and their impact on MT. Finally, the particular challenges and considerations involved in translating literary texts are discussed, setting the stage for later analyses of syntactic and stylistic variation across translation types. Together, these sections establish a comprehensive framework for understanding how MT systems function and how they relate to the specific demands of literary translation.

3.1. Types of Machine Translation

The development of MT systems has evolved significantly over the past decades, moving from rule-based symbolic systems to powerful subsymbolic data-driven neural models (Rothwell, 2023). Each paradigm represents a distinct approach to modeling the translation process, shaped by the availability of linguistic knowledge, computational resources, and training data. This section outlines the main types of MT, rule-based, statistical, and neural, highlighting their underlying methodologies, historical relevance, and current applications. Understanding these foundational systems is essential for contextualizing the capabilities and limitations of modern translation technologies, especially when applied to complex text types such as literary works.

Rule-Based Machine Translation (RBMT) represents the earliest generation of MT systems and was the predominant paradigm throughout the 20th century (Rothwell, 2023). RBMT systems rely on handcrafted linguistic rules and structured bilingual dictionaries, which encode knowledge about morphology, syntax, and lexical equivalence (Moniz and Escartín, 2023). Depending on the design, these systems can follow a direct translation model, a transfer-based model, or an

interlingua-based model as shown in Fig. 3.

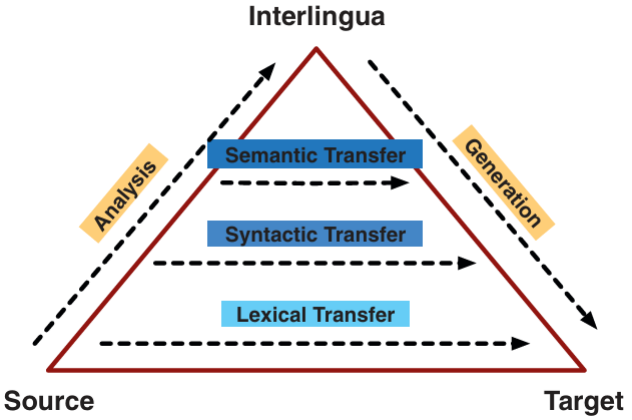

Figure 3.: Vauquois triangle (Koehn, 2020, 11)

Direct translation, or lexical transfer in Fig. 3, uses word-for-word substitution and surface-level rules to adjust the target syntax. In contrast, semantic and syntactic transfer-based approaches involve a deeper analysis of the Source Language (SL) text into abstract representations, which are then mapped to corresponding Target Language (TL) structures. The interlingua approach attempts to represent the input in a language-independent form before generating the output. While no longer the dominant technique, RBMT is still relevant in low-resource scenarios or for closely related language pairs (Poibeau, 2017; Moniz and Escartín, 2023). As a symbolic approach, RBMT is transparent and interpretable but limited in scalability and adaptability (Moorkens, 2025).

Statistical Machine Translation (SMT) emerged in the early 1990s, marking a significant methodological shift toward corpus-based, probabilistic modeling (Moniz and Escartín, 2023). SMT systems learn translation patterns from large aligned bilingual corpora, modeling translation as a process of maximizing the conditional probability of a target sentence t given a source sentence s (Goutte et al., 2009; Koehn, 2010). This process is formalized as a decision problem, where the system selects the target sentence that maximizes the posterior probability $p(t \mid s)$. Applying Bayes' theorem, this can be decomposed into two components as shown in Equation 1 (Goutte et al., 2009, 237).

$$\arg\max p(t \mid s) = \arg\max p(s \mid t) \cdot p(t) \tag{1}$$

In this formulation, $p(s \mid t)$ represents the translation model, which estimates the probability of the source sentence, given a candidate target sentence, while $p(t)$ is the language model, which estimates the grammaticality and fluency of the

target sentence alone. This decomposition allows the two models to be trained independently: the translation model on bilingual parallel corpora, and the language model on large monolingual corpora in the TL. The decoder then searches for the sentence t that balances semantic correspondence and fluency, using both statistical models.

Early SMT systems were word-based, operating on the assumption that individual words are the primary units of translation. However, this approach struggled with multi-word expressions, idiomatic phrases, and many-to-one or one-to-many translation mappings (Koehn, 2010, 2020). As a result, Phrase-Based SMT (PBSMT) became the dominant model, segmenting sentences into sequences of words and modeling their translations probabilistically. More advanced configurations include hierarchical (**?**) and syntax-based SMT (Galley et al., 2006), which attempt to model deeper linguistic structures. Additionally, SMT systems often integrate reordering models to account for structural differences between language pairs (Koehn, 2010). Although SMT marked a substantial improvement over RBMT in terms of fluency and adaptability to domain-specific language, it still operates on surface-level features and often requires PE to reach acceptable quality standards (Poibeau, 2017; Koehn, 2010). While SMT remains symbolically interpretable in theory, its growing model complexity can obscure traceability in practice (Goutte et al., 2009; Moorkens, 2025). SMT systems also demand large quantities of training data and are prone to issues such as error propagation and difficulties in modeling morphologically rich target languages (Scott, 2018; Poibeau, 2017). Despite these limitations, SMT laid the groundwork for many features that are now standard in Computer-Assisted Translation (CAT) tools and continues to inform evaluation practices in the field (Moorkens, 2025).

Neural Machine Translation (NMT) marks a major paradigm shift in MT by applying deep learning methods to model translation as a fully end-to-end process. NMT systems encode input sentences into continuous vector representations and decode them into the TL using neural network architectures (Poibeau, 2017). In contrast to SMT, where translation and language modeling are handled separately, NMT integrates these components into a single trainable system. This integration allows the model to learn linguistic patterns from large-scale parallel corpora, resulting in more fluent and idiomatic output (Moorkens, 2025). Additional advances such as subword-level modeling, multilingual training, and domain adaptation have further improved the scalability and performance of NMT, enabling features such as zero-shot translation and document-level context modeling (Moorkens, 2025). However, NMT systems operate on subsymbolic, high-dimensional vector spaces, making them essentially opaque to human interpreters (Koehn, 2020). While the output is often fluent and natural-sounding, NMT remains susceptible to issues such as hallucinations, omissions, and misinterpretations, especially in cases involving

ambiguity (Moorkens, 2025; Scott, 2018). The high fluency of NMT output can mask such errors, making them harder to detect in PE. These limitations raise ongoing concerns about the application of NMT in sensitive or high-stakes domains and highlight the continued necessity of human revision for complex, stylistically rich content.

The historical progression from RBMT to SMT and finally to NMT encapsulates a broader shift in computational linguistics, from linguistically informed, rule-based engineering to data-driven and subsymbolic learning approaches. RBMT systems, while transparent and interpretable, struggle with scalability and adaptability. SMT improved fluency and domain flexibility through statistical modeling but remained limited in handling deeper linguistic structures and suffered from error propagation and surface-level translation strategies. NMT has significantly enhanced translation quality in terms of fluency, idiomaticity, and generalization by leveraging end-to-end neural architectures and large-scale training data. However, this comes at the expense of interpretability and control, raising concerns about reliability in sensitive or stylistically demanding domains. Each paradigm thus embodies distinct trade-offs between control, scalability, and linguistic nuance, and understanding these differences is essential for assessing their applicability.

3.2. Machine Learning Basics

ML is a field within Artificial Intelligence (AI) that focuses on developing systems capable of improving their performance by learning from data, rather than relying on explicit programming instructions (Murphy, 2012; Campesato, 2020). In the context of MT, ML approaches have become fundamental, shifting the paradigm from hand-coded linguistic rules to data-driven modeling. Especially since the transition from statistical to neural MT, ML methods are central to both system training and ongoing optimization. At its core, ML focuses on automatically identifying patterns in data and using these patterns to make predictions or decisions under uncertainty (Murphy, 2012). Several learning paradigms have shaped MT research and development, most notably supervised, unsupervised, self-supervised, and reinforcement learning. Each of these play a distinct role in MT architectures and their ability to generalize across domains.

Supervised Learning algorithms learn from labeled data. The datasets used for training are labeled, which means they consist of inputs that are paired with their desired output (Campesato, 2020). In a supervised learning setting, we are given a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ consisting of input–output pairs, and the goal is to learn a function $f : X \to Y$ that best approximates the relationship between inputs x and outputs y. Two core task types in supervised learning are classification and regression. In classification, the model predicts discrete labels,

e.g., classifying e-mails as spam or no spam. Common algorithms for classification include decision trees, k-nearest neighbors, Support Vector Machines (SVMs), logistic regression, Naïve Bayes, and ensemble methods such as random forests (Campesato, 2020). Regression tasks involve predicting continuous values, such as estimating housing prices or temperature trends. Supervised learning underpins the training of MT systems on parallel corpora, where each source sentence is labeled with its corresponding target sentence (Koehn, 2020).

Unsupervised Learning operates without labeled data and aims to uncover hidden structures or groupings in the input data (Chapelle et al., 2006). A typical use case is clustering, in which algorithms such as k-means, hierarchical clustering, or expectation maximization attempt to identify coherent groups of instances (Campesato, 2020). These methods help in tasks such as topic modeling, document similarity, and word sense induction. Unsupervised MT attempts to learn translation mappings from multilingual word representations without aligned sentence pairs, often beginning with cross-lingual embeddings derived from unannotated text (Koehn, 2020).

Self-Supervised Learning has gained prominence in deep learning and pre-trained language models, especially in NLP tasks. Self-supervised learning generates labels from the data itself via auxiliary tasks (Koehn, 2020). For example, masked language modeling trains the model to predict missing words based on the surrounding words, the context. Another type of self-supervised learning objectives is to learn to continue a given natural language sequence. In MT, such pre-training allows models to learn linguistic structures from massive unlabeled corpora before being fine-tuned on supervised translation tasks (Koehn, 2020).

Reinforcement Learning (RL) is a distinct learning paradigm in which an agent engages with its environment and learns to take actions that optimize the total reward over time (Koehn, 2020). Rather than learning from labeled examples, the agent learns through feedback provided as rewards or penalties, making RL well-suited to dynamic and sequential decision-making tasks (Campesato, 2020). In NLP, RL has been used in applications such as dialogue systems, text summarization, and interactive MT.

3.3. Neural Architectures for MT

The shift from symbolic to data-driven approaches in MT has been largely driven by advances in neural network architectures. Starting with basic Feed-Forward Neural Networks (FFNNs), the field progressed toward more sophisticated models such as Recurrent Neural Networks (RNNs), encoder-decoder architectures, and the integration of attention mechanisms. These developments culminated in the Transformer architecture, which has become the foundation of most modern MT

systems. This section provides an overview of these key architectures, outlining how they model language, handle sequential data, and enable context-aware translation.

FFNNs form the basic structure of Artificial Neural Networks (ANNs), characterized by information passing in a single direction, from the input layer to the output layer through one or more hidden layers, without any cyclic or feedback connections, as illustrated in Fig. 4. These networks are commonly trained using supervised learning and the backpropagation algorithm, which adjusts the model's weights to minimize the discrepancy between predicted and actual outputs (Jurafsky and Martin, 2025). The multi-layer architecture paired with non-linear activation functions, allows FFNNs to approximate a wide range of functions and extract meaningful patterns from data. Although they lack memory of previous inputs and are thus limited in handling sequential dependencies, FFNNs played a central role in early neural language models and continue to serve as subcomponents in more advanced architectures such as Transformers, where feed-forward layers refine contextual representations after attention mechanisms (Koehn, 2020).

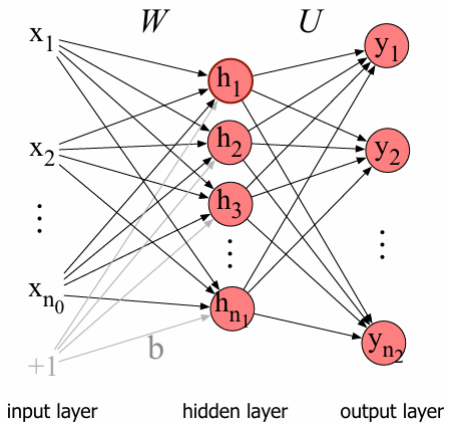

Figure 4.: Feedforward network with one hidden layer (Jurafsky and Martin, 2025, 139)

RNNs extend the architecture of traditional FFNNs by introducing cyclic connections, enabling them to process sequential data through a form of internal memory (Jurafsky and Martin, 2025). This recursive structure makes it possible for the model to draw on information from previous time steps into current computations, making RNNs useful for tasks such as language modeling and MT. Unlike FFNNs, which treat each input independently, RNNs use shared parameters across sequence positions, allowing for consistent modeling of temporal dependencies. However, training RNNs via backpropagation presents challenges, most notably the vanishing gradient problem, in which gradients shrink as they are propagated backward through time, impairing the model's ability to learn long-range depend-

encies (Jurafsky and Martin, 2025). To mitigate this, more advanced variants such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks and Gated Recurrent Units (GRUs) (Cho et al., 2014) have been introduced. LSTMs rely on a system of gates to regulate information flow, retaining relevant content, discarding outdated input, and controlling output, thus preserving gradient flow and allowing the learning of extended dependencies. GRUs offer a simplified alternative with fewer parameters by combining some of these gating functions, often achieving comparable performance with increased training efficiency. While RNN-based models have been largely surpassed by attention-based architectures in modern NLP, their historical significance and utility in sequence modeling remain foundational (Vaswani et al., 2017).

The **encoder-decoder architecture** provides a foundational framework for NMT and other tasks that involve transforming one sequence into another of potentially different length and structure (Jurafsky and Martin, 2025). As the structural backbone of neural MT systems, this design splits the translation process into two components where the encoder converts the input sequence into a context-rich representation, while the decoder constructs the output sequence from this encoded information. Initially, both components were implemented using RNNs. The encoder reads the source sentence and compresses it into a fixed-dimensional hidden state, which the decoder utilizes to produce the translation token by token. This paradigm addressed the challenge of modeling variable-length input-output mappings, especially for structurally divergent language pairs. However, early implementations suffered from bottlenecks when encoding long sequences into a single fixed-size vector (Jurafsky and Martin, 2025). This limitation was later mitigated by architectural innovations such as attention and the Transformer model, which enhance the encoder-decoder framework by enabling the decoder to dynamically focus on different aspects of the input sequence. Despite these advancements, the encoder-decoder architecture remains a central design pattern in contemporary neural MT systems.

The introduction of **attention mechanisms** marked a pivotal advancement in the encoder-decoder architecture, addressing the limitations of relying solely on a single, fixed-size context vector to represent the entire input sequence. First proposed by Bahdanau et al. (2015), attention allows the decoder to dynamically attend to different segments of the source while generating the output. At each decoding step, a context vector is computed as a weighted sum of all hidden states from the encoder, with attention weights reflecting the relevance of each input position to the current output (Koehn, 2020). This mechanism enables more flexible and accurate generation, particularly for longer or structurally complex inputs. By enhancing the stream of information between the encoder and the decoder, the attention mechanism leads to substantial improvements in translation quality

and has become a standard component in modern NMT systems (Vaswani et al., 2017; Koehn, 2020). Beyond the original formulation, various types of attention have since been developed, laying the foundation for later architectures such as the Transformer.

The **Transformer architecture**, introduced by Vaswani et al. (2017), represents a major innovation in NMT by eliminating recurrence in favor of attention-based computation. In contrast to RNN-based models that handle sequences step by step and have difficulty modeling long-range dependencies, Transformers leverage self-attention to capture relationships among all tokens in a sequence simultaneously, allowing for significantly greater parallelization and modeling capacity (Jurafsky and Martin, 2025). The model adopts an encoder-decoder architecture in which both components are built from multiple stacked layers that combine position-wise feed-forward networks with multi-head self-attention as shown in Fig. 5.

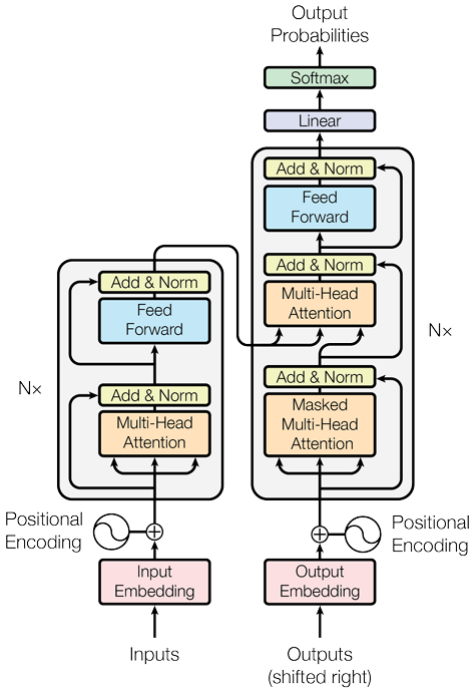

Figure 5.: Transformer architecture (Vaswani et al., 2017, 139)

A key component in the decoder is cross-attention, which enables the model to integrate contextual information from the encoder's representations at each decoding step. To ensure stable training and efficient gradient flow, residual connections and layer normalization are applied after each sublayer. Since Transformers do not naturally recognize the order of sequences, positional encodings are incorporated into the token embeddings to convey the position of each token. This architecture

not only addresses the bottlenecks of earlier encoder-decoder models but also forms the backbone of modern translation systems and large pre-trained language models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), Generative Pre-trained Transformer (GPT) (Radford et al., 2018), or T5 (Raffel et al., 2020), which dominate contemporary NLP applications (Hristea and Caragea, 2022). Its capacity to scale with data and model size while maintaining strong generalization has made the Transformer the de facto standard for a wide range of sequence modeling tasks, including but not limited to MT.

3.4. Large Language Models

LLMs represent the current state of the art in NLP and have significantly reshaped the landscape of MT. Built on the Transformer architecture, LLMs typically employ decoder-only or for MT encoder-decoder structures and are trained on extensive corpora of unlabeled text through self-supervised learning objectives, such as next-token prediction (Jurafsky and Martin, 2025). These models comprise billions of parameters and are capable of generating fluent and coherent text by leveraging contextual information across long input sequences. Through pretraining on massive multilingual and domain-diverse datasets, LLMs acquire a general linguistic competence that can be fine-tuned or prompted to perform a wide range of downstream tasks, including translation. Notable examples include BERT, which is primarily encoder-based and optimized for understanding tasks, and models such as mT5 (Xue et al., 2021) and LLaMA (Touvron et al., 2023), which are text-to-text or generative and can be applied directly to translation scenarios (Hristea and Caragea, 2022). LLMs are particularly powerful in zero-shot and few-shot settings, where they generalize to new languages or tasks with minimal supervision. Their scalability, representational depth, and flexibility make them a central component in contemporary MT systems. However, their generative nature also introduces risks, such as hallucinations, domain sensitivity, and lack of interpretability, which are especially critical in high-stakes contexts such as literary translation (Koehn, 2020). These challenges underscore the need for careful evaluation, domain adaptation, and the continued involvement of human expertise.

3.5. Machine Translation of Literary Texts

In domains with well-defined terminology and repetitive structures, such as technical documentation or user manuals, MT has proven effective. However, its performance in literary texts remains mixed and its application hesitant (Rothwell et al., 2023; Youdale, 2020). In addition, literary translation has been considered incompatible

with MT due to the complex, creative, and culturally embedded nature of literary language (Webster et al., 2020; Hadley et al., 2022). Features such as stylistic nuance, shifts in register, intentional ambiguity, intertextuality, and cultural references present challenges that go beyond the capabilities of early MT systems. As such, literary translators have traditionally emphasized human interpretation, intuition, and creativity as essential qualities that resist automation (Hadley et al., 2022).

However, recent advances in NMT have sparked renewed interest in the applicability of MT to literary texts. While early systems, such as RBMT or SMT, were largely inadequate for literary language, the quality improvements in NMT have led to experimental studies and exploratory applications in this domain (Rothwell et al., 2023; Webster et al., 2020; Hadley et al., 2022).

Research in MT of literature spans several dimensions. It investigates the quality of raw MT output compared to HTs, both through automatic metrics, for example, the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), and human evaluation (Hadley et al., 2022; Fonteyne et al., 2020; Macken et al., 2022). It also explores the use of MT in PE workflows, the cognitive processes during translation or revision, and the potential adaptation of MT systems trained on specific genres or authorial styles (Youdale, 2020; Hadley et al., 2022; Rothwell et al., 2023; Matusov, 2019). Specific challenges include the rendering of metaphor, idiomatic expressions, syntactical structure, stylistic variation, and the preservation of textual cohesion and coherence at the document level.

Studies suggest that MT output often lacks the lexical richness (Toral and Way, 2015), syntactic variation (Toral, 2019; Webster et al., 2020), and literary creativity characteristic (Moorkens et al., 2018) of HTs. Furthermore, machine-translated texts tend to reflect the syntax of the SL more closely, leading to reduced restructuring in the TL, which results in lower syntactic diversity. While grammar is often handled adequately by NMT systems, significant problems remain regarding the accurate rendering of meaning and stylistic nuance (Rothwell et al., 2023). In particular, MT has shown limited success in reproducing cohesive and coherent narratives, especially when it comes to maintaining cohesion across sentence boundaries (Fonteyne et al., 2020; Voigt and Jurafsky, 2012). Despite these limitations, MT has been explored as a productivity-enhancing and time-saving resource, offering a draft from which human translators can work, although this raises questions about output quality, authorial voice, and stylistic loss (Borg, 2022).

A growing body of research also investigates the linguistic fingerprints of MT output, often referred to as machine translationese (de Clercq et al., 2021; Vanmassenhove et al., 2021). This concept parallels the idea of translationese (Gellerstam, 1986) in human translation studies, describing the linguistic characteristics that distinguish translated texts from non-translated originals. As Vanmassenhove et al. (2019) and Webster et al. (2020) have shown, MT systems often produce

output that is strongly oriented toward the structure of the SL and dominated by high-frequency lexical and syntactic patterns.

In sum, while MT of literary texts is a nascent but expanding research area, findings consistently indicate that current NMT systems fall short of capturing the stylistic, interpretive, and creative complexity of literature. Nonetheless, MT may serve a role within controlled workflows, especially when combined with PE or integrated into broader Computer-Assisted Literary Translation (CALT) frameworks. The use of MT in literary translation thus raises important questions about translation quality, authorship, and the evolving role of human creativity in the age of automation. These concerns invite empirical investigation into how the MT output differs from the HT and how PE mediates between the two. These questions are central to the present thesis, which examines the syntactic properties of MT, HT, and PE translations in literary contexts.

4. Post-Editing of Machine Translation

PEMT refers to the process by which human translators revise the raw output generated by MT systems to bring it to an acceptable or publishable level of quality. Within the translation workflow, PE is typically positioned between fully automatic translation and human translation from scratch, forming a hybrid practice that leverages machine-generated content while relying on human judgment and language competence for refinement. As the use of NMT continues to expand across domains, PE has become increasingly central in both industry and academic discourse, including its role in the translation of literary texts (Kolb, 2023; Toral et al., 2018; Moorkens et al., 2018). However, some translators have also expressed skepticism about the practice, a sentiment closely related to the nature of PE itself, which involves "working by correction rather than creation" (Wagner, 1985, 1). This corrective orientation can be perceived as limiting creative agency, particularly in contexts such as literary translation (Moorkens et al., 2018).

4.1. Definition, Challenges, and Post-Editese

The practice of PE is not monolithic but varies considerably in intensity and goals. Traditionally, two levels of PE have been distinguished, namely light and full PE. Their respective definitions follow Nunziatini and Marg (2020). Light PE focuses on minimal intervention to ensure semantic accuracy, often ignoring stylistic or grammatical nuance, and is typically used for internal documents or rapid turnarounds. Full PE, by contrast, seeks to achieve a level of quality comparable to HT, requiring attention to fluency, style, terminology, and contextual appropriateness, as codified in standards such as ISO 18587 (ISO 18587:2017). Nunziatini and Marg (2020) mention the term medium PE as being in the middle of the two types of PE but without further definition. Moreover, they argue for a more nuanced understanding of PE that moves beyond rigid categories, acknowledging that real-world PE often exists along a continuum and less as this binary distinction of full PE or light PE. Industry and research have therefore increasingly embraced tailored PE services, in which quality targets and correction criteria are defined in collaboration with clients based on document type, purpose, and audience

(Nunziatini and Marg, 2020).

A central area of investigation in PE research concerns the different types of effort involved in the task. Krings (2001) identifies three dimensions: temporal, technical, and cognitive effort. Temporal effort is typically measured in terms of editing time per sentence or document. Technical effort refers to the number and type of edits performed, including insertions, deletions, and substitutions that are monitored by cursor movement and keystrokes (Stasimioti and Sosoni, 2020). Cognitive effort, however, remains more elusive and is often studied through eye-tracking, pause analysis, or retrospective interviews (Moorkens et al., 2018).

This negotiation between accepting, rejecting, or rephrasing machine-suggested content forms the basis of translator agency in PE and the engagement of post-editors with the MT output is shaped by their experience, confidence, and the quality of the source (de Almeida and O'Brien, 2010; Borg, 2022). This emphasizes the role of the human being as central in making stylistic and interpretive choices concerning lexical decisions, syntax, and cultural considerations. Empirical studies indicate that professional translators often engage in preferred or preferential changes that reflect their stylistic signature (de Almeida and O'Brien, 2010; Kolb, 2023). While such modifications that go beyond mere error correction can shape the PE text in meaningful ways, the translator's voice may nevertheless be less pronounced than in translations produced from scratch, due to the constraining influence of the pre-existing MT output (Rothwell et al., 2023).

An increasing number of studies examine the linguistic features that characterize PEMT output, commonly referred to as post-editese (Vanmassenhove et al., 2021; de Clercq et al., 2021; Castilho et al., 2019; Castilho and Resende, 2022; Toral, 2019). This phenomenon parallels the more established concept of translationese, describing systematic linguistic traces left by the translation process, such as simplification or explicitation (Gellerstam, 1986; Baker, 1993). In the case of post-editese, these traces often appear more exaggerated than in HT, leading to an exacerbated translationese as Toral (2019) calls it.

Post-editese has been associated with lower lexical diversity, reduced syntactic variation, and closer adherence to SL structures (Toral, 2019; Castilho and Resende, 2022). Neural MT systems tend to reinforce frequent lexical and syntactic patterns while underrepresenting rare or stylistically marked constructions (Vanmassenhove et al., 2019). As a result, post-editors working with such output may be primed to retain the structure and phrasing of the MT suggestions, especially when the output appears superficially fluent. This can lead to what some researchers see as a long-term impoverishment of the TL, particularly when such texts are published without further revision (Toral, 2019; Vanmassenhove et al., 2019; Şahin and Gürses, 2019). The risk of linguistic interference is especially pronounced among novice translators, who may lack the experience to critically assess MT output or feel

constrained by the perceived authority of the machine (Şahin and Gürses, 2019). This makes post-editese a significant object of concern in both the practice and ethics of literary translation workflows.

4.2. Post-Editing in Literary Translation

While MT and PE have long been associated with technical and commercial domains, their extension into the field of literary translation has recently begun to attract increasing academic interest (Rothwell et al., 2023). As literary texts pose a particular challenge to MT systems due to their stylistic complexity, ambiguity, and high degree of cultural and intertextual density, the PE of literary MT output, consequently, requires translators not only to correct errors but also to interpret, reshape, and often recreate meaning with sensitivity to the target audience and literary intent (Castilho and Resende, 2022; Guerberof-Arenas and Toral, 2020).

The process of PE in literary contexts follows the same foundational logic as in other domains. Human translators revise raw MT output to bring it to an acceptable or publishable standard. However, due to the high creative demands of literature, PE in this domain cannot be reduced to surface corrections (Fonteyne et al., 2020). Instead, it becomes a form of collaborative authorship, where the translator negotiates between the machine's suggestions and their own interpretive and stylistic judgments (Rothwell et al., 2023). Although machine-generated drafts can reduce the temporal and technical effort, studies show that the cognitive effort required to edit machine-generated literary translations remains high as translators must often engage deeply with the text to restore cohesion, maintain narrative voice, introduce more creative solutions and resolve stylistic inconsistencies introduced by MT systems (Moorkens et al., 2018; Castilho et al., 2017).

Research into PE of literary texts has demonstrated that, even when MT output is substantially revised, the resulting post-edited texts often retain traces of the machine's influence, also known as post-editese (Castilho and Resende, 2022). These tendencies are typically more pronounced when post-editors are less experienced or when the MT output is of high fluency but low adequacy. The degree to which PE affects the final quality is influenced by variables such as translator experience and genre complexity; figurative or stylistically marked texts often demand deeper intervention than plot-driven ones (Castilho et al., 2019). Although experienced translators may apply preferential changes, the translator's voice often remains less prominent in PE texts, a result attributed to the cognitive anchoring effect of the MT output which continues to fuel skepticism among literary translators (Stasimioti and Sosoni, 2020).

PE has become a key element of modern translation workflows with the rise of NMT, offering efficiency gains in many domains. However, its application to

literary texts raises theoretical and practical challenges that go beyond surface-level correction. Literary PE often requires re-establishing stylistic coherence, cultural nuance, and authorial voice, demanding cognitive and interpretive effort comparable to that of HT from scratch. The phenomenon of post-editese highlights the risk of stylistic flattening and source-language interference. These concerns underline the need for PE guidelines, training, and evaluation frameworks that are sensitive to the demands of literary quality. Ultimately, PE in literary contexts should not be seen as a mechanistic shortcut, but rather as a creative, collaborative process requiring expertise and critical engagement to preserve the integrity of the text and the translator's voice.

5. Evaluating Machine Translation and Post-Editing

This chapter provides an overview of methods used to evaluate the output of MT systems and PE content. It covers both human-centered approaches and automatic metrics, with particular attention to their strengths, limitations, and applicability to literary translation contexts.

5.1. Human Evaluation of Translation Quality

Human evaluation remains the most reliable method to assess the quality of machine-translated texts, as it captures linguistic nuances, contextual suitability, and pragmatic suitability more effectively than automated metrics (Bojar et al., 2017). It is commonly considered the gold standard in both translation studies and computational linguistics. However, it is also inherently subjective, time-consuming, and costly (Popović, 2018; Hiebl and Gromann, 2024). As a result, evaluations are typically performed only on a sample of the output, rather than on entire corpora, limiting coverage and potentially overlooking systematic error patterns (Moorkens, 2025).

Despite these constraints, human evaluation enables nuanced judgments based on two key dimensions: accuracy and fluency. Accuracy, also called adequacy or fidelity, refers to how well the meaning of the source is preserved. Fluency assesses the linguistic well-formedness of the TL output (Fonteyne et al., 2020). Accuracy focuses on semantic completeness, absence of mistranslations, omissions, or additions, while fluency considers grammar, idiomaticity, and stylistic appropriateness.

There are a variety of human evaluation methods. Human ranking compares several MT outputs for a single input, ranking them relatively instead of assigning absolute quality (Bojar et al., 2017). Overall assessments ask evaluators to rate translation quality globally, while Direct Assessment (DA) uses scalar ratings, for example 0–100, and has been shown to achieve high inter-annotator consistency in controlled settings (Bojar et al., 2017). Best-Worst Scaling (BWS) is another time-efficient approach where participants select the best and worst translation from a set, offering a direct, yet intuitive comparison and reducing annotation fatigue (Hiebl and Gromann, 2024).

Manual error annotation is a more diagnostic alternative that allows evaluators to mark and classify translation errors based on predefined categories. This is exemplified by the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014), which provides a hierarchical taxonomy of seven error types concerning different aspects of the text: terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, and design and markup (Lommel et al., 2024). Each category encompasses specific subtypes, e.g., mistranslation, inconsistent use of terminology, or spelling. In addition, severity levels are assigned from neutral and minor to major and critical. MQM has gained traction in MT research for its granularity and transparency, despite being resource-intensive and prone to inter-annotator variability. Sampling remains a necessary compromise in such settings, but it also constrains the representativeness of the evaluation and the generalizability of its conclusions. The combination of manual and automatic processes involved in error annotation workflows can be seen in Fig. 6.

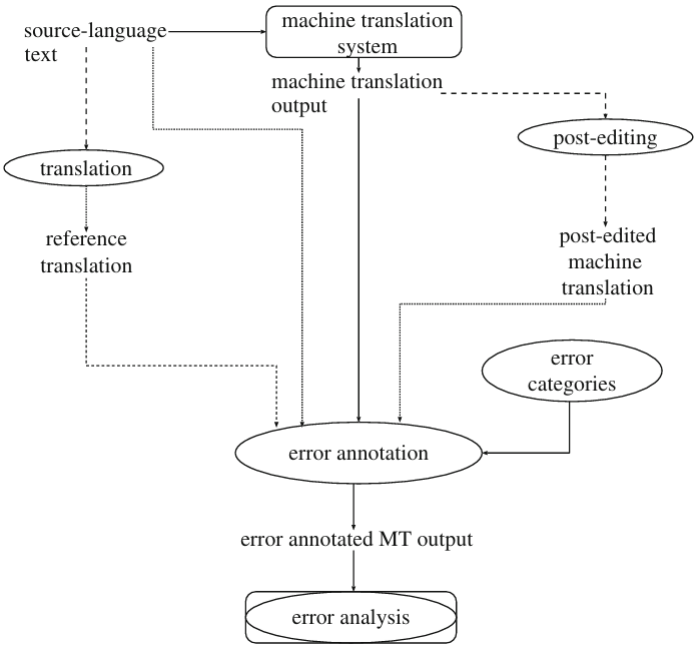

Figure 6.: Manual error annotation with automatic processes in rectangles and manual processes in ellipses (Popović, 2018, 132)

Finally, task-based evaluations and "fitness for purpose" (Way, 2013, 2) assessments offer a functionalist perspective aligned with translation studies. Rather than aiming for semantic or structural equivalence, these approaches judge translation quality in terms of the target audience, communicative function, and usability by asking whether the translation is sufficient for its intended purpose (Liu et al.,

2024). This allows quality assessment to shift from an idealized notion of perfection to a context-aware, goal-oriented paradigm.

5.2. Traditional MT Assessment

Automatic evaluation metrics are widely used in both academic research and industry due to their efficiency, scalability, and consistency. These metrics typically assess the quality of machine-generated translations by comparing them to one or more human-produced reference translations. While they do not capture all linguistic nuances, they serve as practical proxies for translation quality and enable system development and benchmarking at scale (Moorkens, 2025). Their underlying assumption is that the quality of an MT output can be approximated by measuring its similarity to a reference.

One of the earliest and most influential metrics is the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002). BLEU measures the degree of n-gram overlap between a candidate translation and reference translations, using precision to calculate how many n-grams in the hypothesis also occur in the reference(s). It imposes a brevity penalty to prevent translations from being excessively short. Despite its historical significance and widespread adoption, BLEU has faced criticism for its insensitivity to paraphrasing, meaning preservation, and syntactic variation, issues that particularly relevant in literary or stylistically rich texts.

The National Institute of Standards and Technology (NIST) metric (Doddington, 2002) builds on BLEU by assigning higher weights to less frequent, and thus more informative n-grams. Instead of treating all matches equally, NIST computes information weights from n-gram frequencies in the reference translations, so that matches of rarer n-grams contribute more to the score than common ones. Like BLEU, NIST evaluates n-gram overlap and includes a penalty for length discrepancies, but the brevity penalty was modified to reduce the impact of minor length variations while still discouraging short translations. Neverthelless, as a lexical n-gram-based metric, NIST continues to face limitations in capturing deeper semantic and syntactic equivalence.

Translation Edit Rate (TER) (Snover et al., 2006) is a distance-based metric that quantifies the number of modifications needed to convert a MT into a reference translation. These modifications encompass insertions, deletions, substitutions, and shifts. TER is particularly relevant in PE contexts as it can reflect the technical effort involved in correcting machine outputs. A lower TER indicates a closer match and, by extension, a potentially lower PE workload.

Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee and Lavie, 2005) was introduced to address some of BLEU's shortcomings. It computes unigram-level matches using flexible mechanisms such as stemming, syn-

onymy matching, and paraphrase recognition. Unlike BLEU, METEOR calculates both precision and recall and penalizes translations in which the matched words do not appear in the same order, a mechanism called the fragmentation penalty. This increases its sensitivity to word order changes and meaning retention, enabling it to align more closely with human judgments, particularly at the segment level.

ChrF (Popović, 2015) is an automatic MT evaluation metric based on character-level n-gram F-scores. Unlike traditional word-based metrics such as BLEU or METEOR, chrF operates at the level of character n-grams, making it particularly suited for morphologically rich or low-resource languages where surface-level word matching may be unreliable. The metric computes both precision and recall of character n-grams between the machine-generated translation and the reference, combining them into a single F-score. ChrF offers several advantages, as it is simple to compute, language-independent, tokenization-free, and does not rely on external linguistic resources. These features make it particularly useful in multilingual and noisy data scenarios. It is also sensitive to finer-grained differences such as inflectional morphology or small orthographic changes, which are often overlooked by word-based metrics.

Crosslingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020) represents a shift toward neural, learned evaluation metrics. Rather than relying on surface similarity, COMET uses contextualized multilingual language models to encode the source, hypothesis, and optionally a reference into contextual embeddings. A regression model then predicts a quality score that approximates human judgment. This approach allows COMET to model deeper semantic relationships and generalize across different languages and domains. It has been shown to outperform traditional metrics in correlating with human direct assessment scores. A recent extension of this framework is xCOMET (Guerreiro et al., 2024), whose training includes human-annotated MQM data to more closely align with fine-grained quality judgments. By incorporating both sentence-level and span-level supervision during training, xCOMET improves the interpretability of its scores and enables more diagnostic insights into translation quality. This makes it particularly useful for evaluations where explainability and error categorization are essential.

In general, while traditional surface-based metrics, such as BLEU, remain widely used for benchmarking, the MT community increasingly recognizes the need for linguistically informed and reference-aware or even reference-independent metrics that better reflect human perceptions of quality, particularly in contexts that require semantic accuracy and stylistic nuance (Rei et al., 2020).

5.3. Syntactic Evaluation

While traditional MT evaluation metrics such as BLEU or METEOR focus on surface-level correspondence between translations and reference texts, they often fail to account for deeper syntactic or stylistic characteristics that are especially crucial in literary translation. Beyond surface-level correctness, translation quality also involves syntactic fidelity and stylistic adequacy. Studies have shown that MT systems, especially in their neural implementations, tend to produce outputs that closely mirror the syntactic structure of the ST, often more so than HTs (Webster et al., 2020; Volkart and Bouillon, 2024). This phenomenon can be examined through POS patterns, clause length distributions, or dependency structures. PEMTs often reflect these tendencies as well, with shorter average sentence lengths and structural proximity to the source. These characteristics are collectively referred to in the literature as indicators of post-editese or translationese.

In order to better understand the structural and grammatical transformations that occur across translation types, namely HT, PEMT, and MT, a syntactic and stylistic evaluation is necessary. This chapter focuses on three linguistically motivated metrics proposed by Vanroy et al. (2021) that quantify syntactic equivalence at the sentence level: Word Label Changes, Syntactically Aware Cross (SACr), and Aligned Syntactic Tree Edit Distance (ASTrED). These measures rely on Universal Dependencies (UD) annotations and word alignment to capture different facets of syntactic variation.

5.3.1. Word Alignment

In order to compute the syntactically motivated evaluation metrics introduced in the following sections, word-level alignments for the parallel data are required. For this purpose, SimAlign (Jalili Sabet et al., 2020), a state-of-the-art alignment framework, was employed. It produces high-quality word alignments based on multilingual embeddings, without requiring any parallel data. Departing from traditional statistical alignment approaches such as GIZA++ (Och and Ney, 2003), fastalign (Dyer et al., 2013), or efmaral/eflomal (Östling and Tiedemann, 2016), which depend on large parallel corpora, SimAlign achieves competitive or, for a few languages, superior results using only monolingual embeddings. SimAlign constructs an embedding-based similarity matrix between words of the aligned and tokenized source and target sentences before extracting alignments using different strategies, each trading off between precision and recall. The Intersection (Inter) strategy selects only those alignments that occur in both forward and backward directions which minimizes noise but sacrifices recall. The Maximum Weight Matching Forward (MWMF) strategy models the alignment process as a matching problem between two sets of words, representing source and target tokens. It builds

a graph where each word from the source sentence is connected to each word of the target sentence, with edge weights determined by embedding-based similarity scores. The algorithm then finds an optimal one-to-one assignment that maximizes the total similarity. The Argmax method aligns a source and target word if they are each other's most similar counterpart in the similarity matrix, identifying local alignment pairs. Building on this, the Itermax strategy applies Argmax iteratively, updating the similarity matrix after each step to encourage new alignments. Unlike MWMF, Itermax does not guarantee a globally optimal solution bun can generate many-to-many alignments and often achieves higher coverage.

5.3.2. Word Label Changes

One way to evaluate the syntactic differences between a source sentence and its translation is to analyze how the grammatical roles of words shift between the two. This type of syntactic comparison can be quantified using Word Label Changes (WLC), a metric discussed by Vanroy et al. (2021). This method compares dependency labels, i.e., the annotations that describe the syntactic role of each word in a sentence, such as subject, object, or root, across aligned word pairs in the source and target texts.

In practical terms, a label change occurs whenever a word in the source sentence fulfills one syntactic function, e.g., subject, and its translated equivalent fulfills a different function, e.g., object, in the target sentence. This often happens when the sentence structure changes during translation, for instance, when an active sentence is transformed into a passive one. Such changes are not errors, they reflect the translator's effort to adapt the sentence to the language and style standards of the target language.

To illustrate how label changes are calculated in syntactic evaluation, consider Example (1), adopted from Vanroy et al. (2021). The English sentence *I saw him* is translated into Dutch as *Hij werd door mij gezien* (He was seen by me). This translation involves a shift from active to passive voice, resulting in changes in the grammatical functions of several constituents. The corresponding dependency labels for each token in the source and target sentences are listed in (a) and (b), respectively, using UD notation. In (c), the alignments of the words between the source and target tokens are indicated, mapping each source word to its aligned word in the translation, which is further visualized in Fig. 7.

(1)	(a)	I	saw	him		
		nsubj	root	obj		
	(b)	Hij	werd	door	mij	gezien
		He	was	by	me	seen
		nsubj	aux	case	obl	root

(c) 0-2 0-3 1-1 1-4 2-0

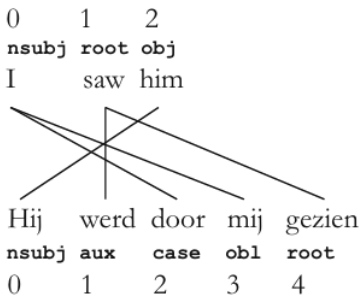

Figure 7.: Visualization of the word alignment (Vanroy et al., 2021, 273)

Table 1 summarizes the syntactic label changes that occur as a result of these alignments. Each row compares a source–target word pair along with their respective dependency labels. A change is counted when the dependency labels differ between aligned tokens. For instance, the subject *I* in the source is aligned with the preposition *door* (by) and the pronoun *mij* (me) in the target, both of which carry different syntactic labels, case as in case marking and obl as in oblique nominal. This results in two label changes. The verb *saw*, marked as the root in the source, aligns with both the auxiliary verb *werd* (was) and the root verb *gezien* (seen) in the target, of which only the latter retains the same syntactic role. Overall, four out of five alignments involve label changes, resulting in a normalized label change score of 0.8.

Table 1.: Example (1) with label changes between source and target sentences (Vanroy et al., 2021, 273)

Source (label)	**Target (label)**	**Change**
'I' (`nsubj`)	'door' (`case`)	1
'I' (`nsubj`)	'mij' (`obl`)	1
'saw' (`root`)	'werd' (`aux`)	1
'saw' (`root`)	'gezien' (`root`)	0
'him' (`obj`)	'Hij' (`nsubj`)	1
	Total:	**4 (normalised: 4/5 = 0.8)**

Given a set A of aligned source and target syntactic label pairs for a sentence and its translation, the total number of label changes L is defined as shown in Equation 2.

$$L = \#\{(s,t) \in A : s \neq t\} \tag{2}$$

Here, A denotes the set of alignment pairs (s,t), where s is the syntactic label in the source sentence and t is the syntactic label in the target sentence. It should be noted that in cases where a label aligns with multiple counterparts, e.g., many-to-one, one-to-many, or many-to-many alignments, each such pair is counted separately.

Label changes are especially useful when analyzing structural fidelity in translation, in other words how closely the grammatical structure of the source sentence is mirrored in the target sentence. In literary or stylistically complex texts, translators often depart from literal grammatical structures to preserve idiomaticity or stylistic nuance. By measuring label changes, researchers can gain insight into how and where such syntactic shifts occur, making this metric a valuable tool in comparative translation studies.

5.3.3. Syntactically Aware Cross

SACr (Vanroy et al., 2021) quantifies the degree of reordering of word groups in translation and builds on an earlier metric known as seq_cross (Vanroy et al., 2019). Unlike its predecessor, SACr uses linguistic structure to validate the word groups being compared. Specifically, it relies on dependency trees to ensure that each word group forms a valid subtree, meaning that all the words within a group share direct parent-child relationships. This linguistic grounding enables a more meaningful interpretation of word reordering, particularly in cases where surface changes in word order do not necessarily reflect syntactic reorganization.

To calculate SACr, consecutive words are grouped based on alignment links and checked for syntactic validity. If a group does not form a valid subtree, it is split into smaller syntactically valid subgroups. The number of alignment crossing links between these groups is then counted and normalized by the total number of alignments. A higher SACr score indicates greater reordering, which may correlate with more syntactic restructuring and, potentially, increased cognitive effort during PE or reading.

Fig. 8 demonstrates how SACr analyzes the English source and its Dutch translation by comparing the positions and dependency relations of aligned word groups. The SACr value is obtained by dividing the number of crossing links (3) by the total number of alignments (9), yielding a normalized score of 0.33. This metric is particularly effective in capturing phrasal reordering across languages with differing syntactic preferences and can provide insights into the syntactic creativity or literalness of a translation.

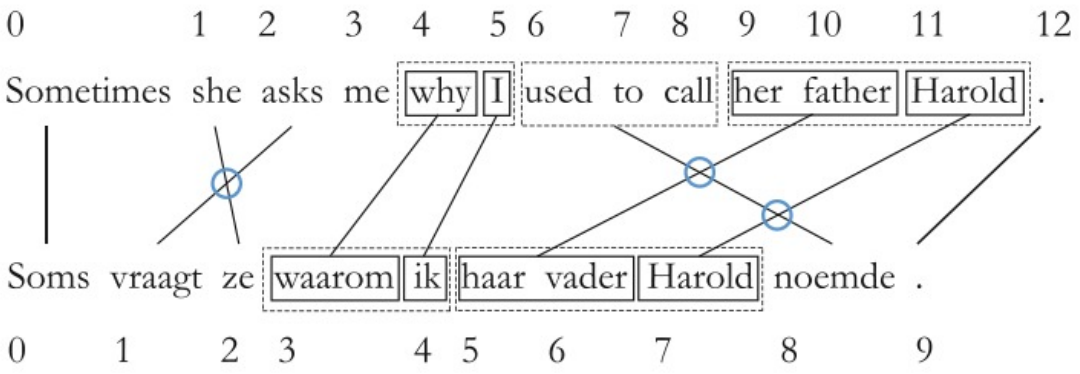

Figure 8.: Illustration of SACr showing initial seq_cross groups in dotted boxes and new linguistically motivated SACr groups in solid boxes based on dependency relationships. SACr value of 3/9 = 0.33 (Vanroy et al., 2021, 270)

5.3.4. Aligned Syntactic Tree Edit Distance

The ASTrED (Vanroy et al., 2021) metric provides a structural comparison of syntactic trees by quantifying how much transformation is needed to convert the dependency tree of a source sentence into that of the corresponding target sentence. The method starts by parsing both source and target sentences into dependency trees, then mapping aligned tokens between them. To align the syntactic structures, labels of the aligned tokens are adjusted where necessary to reflect consistency.

Using the aligned and updated trees, the metric calculates the Tree Edit Distance (TED) required to transform one tree into the other. This score is then normalized by the average number of words in the source and target sentences. ASTrED is able to capture deeper structural differences than surface-level metrics such as label changes or SACr, and it also accommodates cases of null alignments and varying sentence lengths, which are common in translations across languages with different syntactic norms. Example (2) shows the alignment of the English sentence *Does he believe in love?* and its Dutch translation.

(2) (a) Does he believe in love ?
aux nsubj root case obl punct

(b) Gelooft hij in de liefde ?
Believes he in the love ?
root nsubj case det obl punct

(c) 0-0 1-1 2-0 3-2 4-3 4-4 5-5

Visualizations of the respective dependency trees in both languages are shown in Fig. 9 and Fig. 10.

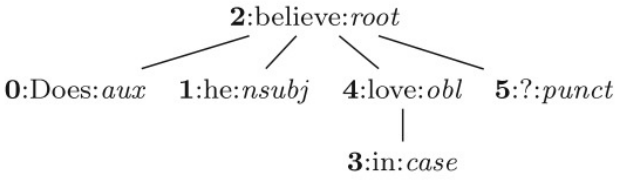

Figure 9.: Dependency tree of the source sentence of Example (2) (Vanroy et al., 2021, 275)

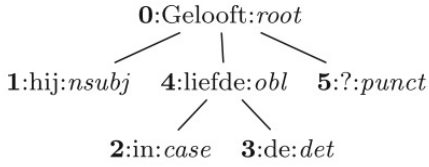

Figure 10.: Dependency tree of the target sentence of Example (2) (Vanroy et al., 2021, 275)

In Example (3), both trees are aligned at the level of corresponding subtrees. After the label adjustments, the final edits required to convert one tree into the other are shown in Fig. 11.

(3)
- `aux:root|root:root` (does:gelooft | believe:gelooft)
- `nsubj:nsubj` (he:hij)
- `case:case` (in:in)
- `obl:det,obl` (love:de,liefde)
- `punct:punct` (?:?)

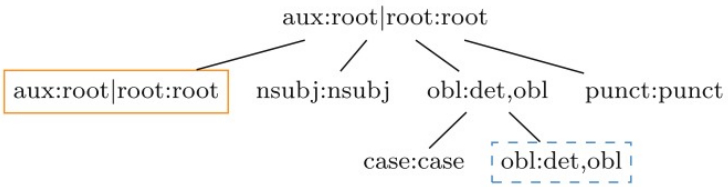

Figure 11.: Edited tree with two necessary edits. Orange signifies deletion and blue signifies insertion (Vanroy et al., 2021, 277)

ASTrED enables detailed structural comparison and reveals translation-specific syntactic transformations, offering complementary insights alongside surface-level metrics. When applied to translations of literary texts, it helps uncover subtle, yet meaningful structural differences across MT, PEMT, and HT outputs.

As summarized in Table 2, the metrics introduced here operationalize syntactic equivalence along three main dimensions: differences in dependency label of aligned words, differences in word group order, and differences in overall syntactic structure. Label changes address the first aspect, SACr captures the second by focusing on reordering phenomena, and ASTrED extends the comparison to full dependency trees. Together, these measures provide a layered perspective on syntactic diversity across MT, PEMT, and HT. The following chapters build on this framework with related work within existing research and a detailed account of the methodology applied to the literary dataset.

Table 2.: Overview of the metrics introduced in this chapter (Vanroy et al., 2021, 277)

Metric	**Captures**	**Normalisation by**
Label Changes	Changes in dependency labels in the surface form based on word alignment	Number of alignments
SACr	Reordering of linguistically motivated groups by measuring crossing links	Number of alignments
ASTrED	Structural difference between the source and target dependency trees while also taking word alignment into account	Average number of source and target words

6. Related Work

This thesis compares HT, MT and PE of literary texts. Previous work has shown that PE remains syntactically more similar to the ST (Toral, 2019). This leads to the assumption that this does not differ when observing literary translations.

Volkart and Bouillon (2024) compared HT and PEMT to determine the effect of MT on the lexical and syntactic level of generated translations. Three metrics were used concerning the syntactic equivalence: label changes to capture differences in dependency labels for word-aligned pairs, SACr to gain insight on the degree of reordering of word sequences, and ASTrED which compares dependency trees with the help of word alignments to observe structural differences. This results in the observation that PEMT plays a role in making the raw MT output syntactically more diverse through word reordering and changing the dependency labels. Still, PEMT is closer to the source when compared to HT. The data used for the comparative analysis consists of rather technical texts, documents from the European Investment Bank, translations from a sports organization, and texts from an insurance company. In contrast, the objective of this thesis is to have a set of literary texts as a basis.

Shaitarova et al. (2023) apply NMT systems on a diverse set of parallel corpora to translate into German. Lexical, syntactic, and morphological features of both, the MT output and the human reference, are compared. ASTrED was applied to capture the syntactic differences and equivalence between source and target text. The results demonstrate that MTs do not show the same degree of syntactic creativity as HT. In fact, the sentence structure of the machine generated translations stays close to the syntax of the ST. Although the data set includes one literal novel by Jane Eyre, the empirical work does not focus on literary translations, which is the main point of interest in this thesis.

Ahrenberg (2017) compared a Swedish HT and a Swedish MT of a British opinion article to demonstrate the limitations in MT, albeit in a small setting. Statistics, such as sentence structure and word order, reveal that the MT is more similar to the source than the HT. Furthermore, in the HT translation procedures such as sentence splitting or explication are detected which the MT does not present. This leads to an unpublishable product for the MT. In combination with PE this may change but in order to conclude if the target text is then on the same level as an HT, PE texts should be included in a similar study.

Castilho and Resende (2022) inspected the specific set of features of a PE text that set it apart from a previous HT, also known as the post-editese phenomenon.

The study was based on two English literary texts that were translated into Brazilian Portuguese via Google Translate. They used the translationese features, simplification, explicitation, and convergence as a basis to identify similar patterns in the PE texts. Although, not all assumed features associated with post-editese were confirmed. Especially convergence, which expresses the variation in individual texts, tends to be lower for PEs than their respective HTs. But this cannot be generalized over the whole genre of literary texts as there are differences in the various sub-genres. Still, they mention the priming effect as a possible cause for similar features between the MT output and the PE product. Further research in other sub-genres, with different PE tools, and a more balanced dataset with more than one ST and MT output may help to support this observation.

Castilho et al. (2019) investigate the presence of post-editese features in PEMT compared to translationese in HT. They examine the influence of different factors, such as light and full PE, news and literary domain, and translator proficiency, where they compare students and professional translators. Using linguistic features such as lexical richness, lexical density, and sentence length, they found that light PE introduces more post-editese features than full PE, with significant domain-dependent variations. The findings also show that PE texts are closer to MT outputs than HT in terms of syntactic and lexical patterns. This study highlights the influence of human interference and domain characteristics, but leaves open questions regarding the application of these findings to syntactic diversity in literary texts, which this thesis aims to address.

7. Methodology

The objective of this thesis is to investigate the syntactic diversity of different types of translation for literary texts, including MT, PEMT, and HT. To this end, a corpus of Brazilian Portuguese short stories by two authors was collected, each accompanied by two translation modalities into German, HT, or in one case PEMT, and MT. The analysis proceeded in several stages. First, the texts were preprocessed and sentence-aligned to ensure comparability. The word alignment was then applied to link tokens across translation pairs, providing the basis for syntactic comparisons. After this, a range of syntactic diversity measures was computed, including dependency label changes, POS tag changes, word and sequence crossings, SACr, and ASTrED. Finally, selected examples were manually inspected to qualitatively illustrate and contextualize the observed patterns. This chapter describes each of these steps in detail, beginning with the composition and preprocessing of the dataset, followed by the alignment procedures, the application of syntactic metrics, and the setup of the qualitative inspection.

7.1. Dataset

This study is based on 24 literary STs originally written in Brazilian Portuguese, comprising 22 short stories by Lima Barreto and 2 by Machado de Assis. Each text has a corresponding MT, and most of them also have an HT. One text, an excerpt from Lima Barreto's *Diário do Hospício*, is available as a PEMT and not as an HT. Table 3 shows the whole extent of the used dataset with the total character, words, and sentence count along with the average TTR and sentence length. It highlights some initial differences between modalities. The MT corpus is identical in sentence count to the ST, while the HT corpus is slightly shorter. The single PEMT sample is considerably smaller in scope, as it consists of a single text. In terms of lexical variety, the TTR is highest for the HT and lowest for the PEMT, indicating that HTs employ richer vocabulary. While lexical diversity is not the focus of this thesis, it may still influence certain syntactic measures such as dependency label changes. Sentence length is broadly comparable across modalities, though the PEMT sample shows a tendency toward longer sentences. This reflects the fact that PEMT is represented by a single text in the dataset and may therefore not be fully representative of the modality as a whole. Nevertheless, the values provide

a useful point of comparison for situating PEMT within the broader spectrum of translation types.

Table 3.: Corpus statistics for the dataset

Metric	ST	MT	PEMT	HT
Number of sentences	1,055	1,055	88	967
Total words	16,467	15,715	1,827	14,720
Characters	89,222	101,964	11,345	94,672
TTR	0.533	0.592	0.462	0.606
Mean sentence length	17.55	16.75	20.76	17.24

Some of the STs and their corresponding translations include editorial footnotes. Since these are not part of the narrative texts itself and vary in presence and frequency across translations, they were excluded from the syntactic diversity analysis. In ten cases, the number of footnotes in the HT differs from the ST, while in others no footnotes are present at all. Including them would artificially increase sentence and word counts, and thus inflate syntactic diversity measures, making a direct comparison across modalities unreliable. For this reason, the analysis focuses only on the main body of the texts without footnotes, even though this gives a restricted view of the full HT in some cases. This restriction was necessary to ensure comparability across MT, PEMT, and HT versions. The following subsections provide an overview of the dataset, including representative examples from each author and statistical summaries of the selected texts. Full corpus statistics are available in Appendix A.1.

7.1.1. Lima Barreto Texts

The bulk of the dataset consists of 22 short stories by Lima Barreto (1881-1922), a prominent early 20th-century Brazilian author known for his socially critical and linguistically rich prose (Aidoo et al., 2013). For each of these texts, a corresponding MT was generated using DeepL's NMT system,[1] and 21 of them also have HTs produced as part of a collaborative translation project commemorating the 2022 Bicentennial of Brazil's Independence (Ugalde, 2023). One additional text, an excerpt from *Diário do Hospício*, is available in a PEMT version. This text was produced as part of a research study investigating literary PE in a real-world context. As Kolb (2023) describes, a short narrative by Lima Barreto was first translated into German using DeepL's NMT system. A professional literary translator, who was also co-editor of the anthology where the text was published, then post-edited

[1] `https://www.deepl.com/de/translator`

the MT draft. This entire process was captured using keylogging software to allow for an analysis of the PE workflow and the types of edits made. Table 4 presents an excerpt from this PEMT text, showing the Portuguese source, the MT output, and the PE version.

Table 4.: Source paragraph and its two translations of Barreto's *O Pavilhão e a Pinel*

Portuguese source	**German MT**	**German PE**
Estou no Hospício ou, melhor, em várias dependências dele, desde o dia 25 do mês passado. Estive no pavilhão de observações, que é a pior etapa de quem, como eu, entra para aqui pelas mãos da polícia. Tiram-nos a roupa que trazemos e dão-nos uma outra, só capaz de cobrir a nudez, e nem chinelos ou tamancos nos dão. Da outra vez que lá estive me deram essa peça do vestuário que me é hoje indispensável. Desta vez, não. O enfermeiro antigo era humano e bom; o atual é um português (o outro o era) arrogante, com uma fisionomia bragantina e presumida.	Ich bin seit dem 25. des letzten Monats im Hospiz, oder vielmehr in verschiedenen Teilen davon. Ich war auf der Beobachtungsstation, die die schlimmste Stufe für diejenigen ist, die wie ich hier in die Hände der Polizei geraten. Sie nehmen dir die Kleidung weg, die du trägst, und geben dir eine andere, die nur deine Blöße bedeckt, und sie geben dir nicht einmal Hausschuhe oder Clogs. Das andere Mal, als ich dort war, haben sie mir dieses Kleidungsstück geschenkt, das jetzt für mich unverzichtbar ist. Diesmal nicht. Die alte Krankenschwester war menschlich und gut; die jetzige ist eine arrogante Portugiesin (die andere war es), mit einer prahlerischen und anmaßenden Physiognomie.	Seit dem 25. des vergangenen Monats bin ich in der Irrenanstalt, genauer, in ihren verschiedenen Abteilungen. Zunächst war ich im Beobachtungspavillon, der schlimmsten Station für einen, der wie ich von der Polizei eingeliefert wurde. Man nimmt uns, was wir am Leibe tragen, und gibt uns ein Hemd, das kaum die Blöße bedeckt, nicht einmal Hausschuhe oder Pantoffeln bekommen wir. Das letzte Mal gaben sie mir noch welche, sie sind mir unverzichtbar geworden. Diesmal nicht. Der Krankenpfleger damals war menschlich und gut; der jetzige ist (wie schon der vorige) Portugiese, aber überheblich, mit den selbstgefälligen Gesichtszügen aus dem Hause Braganza.

General statistics for this subset, such as sentence count, word count, character length, and TTR for every text, are summarized in Appendix A.1. The statistics for this specific text can be seen in Table 5. It is particularly relevant because it is the only case where a PEMT is available instead of an HT, making it central

for evaluating how PE compares to raw MT and the ST. As the values indicate, MT and PEMT remain close to the ST in terms of length, but there is a gradual increase in lexical variety, see TTR, from ST to MT to PEMT, accompanied by a slight rise in word count. This progression contrasts with the overall corpus trend where PEMT shows the lowest TTR. The difference can be explained by the fact that PEMT in this study is represented by a single text rather than an average across multiple texts, making it less representative of a general tendency.

Table 5.: Corpus statistics for Barreto Text 1: *O Pavilhão e a Pinel*

Metric	**ST**	**MT**	**PEMT**
Number of sentences	88	88	88
Total words	1,803	1,804	1,827
Unique lemmata	598	625	700
Characters	9,501	11,027	11,345
TTR	0.419	0.431	0.462
Mean sentence length	20.49	20.50	20.76
Average length difference	–	2.64	2.86

To complement this, Table 6 presents the statistics for another Barreto text, which follows the standard setup of ST, MT, and HT. Here, HT displays the highest lexical variety and longest sentences.

Table 6.: Corpus statistics for Barreto Text 4: *A lei*

Metric	**ST**	**MT**	**HT**
Number of sentences	13	13	13
Total words	288	296	302
Unique lemmata	141	144	156
Characters	1,575	1,892	2,010
TTR	0.531	0.588	0.616
Mean sentence length	22.15	22.77	23.23
Average length difference	–	2.92	4.92

These values are in line with the overall corpus averages reported in Table 3, where HT generally displays higher lexical richness and slightly longer sentences on average than MT. This pattern also reflects previous findings that HTs tend to be lexically and syntactically more diverse, while MT output remains more conservative and closer to the source (Vanmassenhove et al., 2021; Toral, 2019). Including this second example highlights the contrast between a regular HT case

and the unique PEMT case discussed earlier, therefore providing a fuller picture of the dataset's characteristics. These statistics are relevant for the subsequent syntactic analysis, as differences in lexical variety and sentence length may shape the extent of syntactic restructuring across modalities.

7.1.2. Machado de Assis Texts

In addition to the works by Barreto, the dataset contains two texts by Joaquim Maria Machado de Assis (1839-1908), a canonical figure in Brazilian literature whose writing is often characterized by irony, subtlety, and syntactic complexity (Graham, 1999). Each Machado text is available in an MT and an HT version. These two texts serve to broaden the stylistic and linguistic scope of the corpus. A representative excerpt from one Machado text, *O Dicionário*, can be seen in Table 8 along with comparative statistics in Table 7. As shown in Table 7, the MT version remains close to the ST, in some cases even showing a slight decrease compared to the source. By contrast, the HT surpasses both ST and MT in every aspect as it contains more words, more characters, and a higher number of unique lemmata. This also results in the highest TTR, reflecting greater lexical variety, and slightly longer average sentences. Taken together, these values support the initial assumptions of this thesis that HTs tend to introduce more restructuring and greater lexical diversity, while MT remains closer to the ST. The Machado texts therefore provide further evidence for the overall corpus tendency that HT diverges more syntactically from the ST than MT outputs.

Table 7.: Corpus statistics for Machado Text 1: *O Dicionário*

Metric	**ST**	**MT**	**HT**
Number of sentences	39	39	39
Total words	808	795	822
Unique lemmata	317	327	372
Characters	4,474	5,130	5,392
TTR	0.481	0.532	0.563
Mean sentence length	20.72	20.38	21.08
Average length difference	–	2.05	3.00

Table 8.: Source paragraph and its two translations of Machado's *O Dicionário*

Portuguese source	**German MT**	**German HT**
ERA UMA VEZ um tanoeiro, demagogo, chamado Bernardino, o qual em cosmografia professava a opinião de que este mundo é um imenso tonel de marmelada, e em política pedia o trono para a multidão. Em mim, bradou ele, podeis ver a multidão coroada. Eu sou vós, vós sois eu. O segundo foi declarar que, para maior lustre da pessoa e do cargo, passava a chamar-se, em vez de Bernardino, Bernardão. Particularmente encomendou uma genealogia a um grande doutor dessas matérias, que em pouco mais de uma hora o entroncou a um tal ou qual general romano do século IV, Bernardus Tanoarius.	Es war einmal ein demagogischer Küfer namens Bernardino, der in der Kosmographie die Meinung vertrat, dass diese Welt ein riesiger Marmeladentonne sei, und in der Politik forderte er den Thron für das Volk. In mir, rief er, könnt ihr die gekrönte Menge sehen. Ich bin ihr, ihr seid ich. Als zweites erklärte er, dass er zu größerem Glanz seiner Person und seines Amtes fortan statt Bernardino Bernardão heißen werde. Er gab eigens einem großen Gelehrten dieser Materie den Auftrag, eine Genealogie zu erstellen, die ihn in etwas mehr als einer Stunde mit einem bestimmten römischen General aus dem 4. Jahrhundert, Bernardus Tanoarius, verband.	Es war einmal ein Böttcher und Demagoge namens Bernardino, der in kosmografischen Belangen die Ansicht vertrat, die Welt sei ein riesiges Fass Marmelade, und in politischen, dass der Thron der Masse gebühre. In mir, donnerte er, seht ihr die gekrönte Masse! Ich bin ihr, ihr seid ich. Als Zweites verkündete er die Änderung seines Namens, zu höherem Glanz von Amt und Person, von Bernardino zu Bernardonus. Zudem gab er bei einem großen Gelehrten dieses Fachs einen Stammbaum in Auftrag, der ihn binnen einer Stunde zum Nachfahren eines römischen Generals des vierten Jahrhunderts machte, eines gewissen Bernardus Botticus.

7.2. Experimental Setup

While the STs, the HTs, and the PEMT were available from the start, the respective MTs were created using the NMT system DeepL. Pre-processing and basic linguistic annotation were conducted using the Stanza NLP toolkit, which was applied to both the Portuguese and German texts for sentence segmentation, tokenization, POS tagging, dependency parsing, and constituency parsing. Although syntactic comparison between the translations is the primary focus of this thesis, the Portuguese STs played an essential role for alignment and baseline statistics. As noted

in Section 7.1, footnotes were excluded from the texts to ensure comparability across modalities.

Sentence segmentation proved to be inconsistent across the two language models in Stanza. The German model, for instance, frequently split sentences at semicolons, while the Portuguese model did not. This discrepancy, alongside translator choices, required manual sentence adjustment, realignment, or, in some cases, sentence deletion to obtain machine-readable parallel data. These adjustments in the texts were performed while ensuring that they do not change the content or structure of the texts but focused on aligning them for further processing. The sentences that could not be adjusted for other reasons were removed. Redundant or problematic elements, such as direct speech markers were removed uniformly across all texts for cleaner processing, e.g., tokenization or sentence segmentation, and cleaner metric evaluation. In cases where punctuation affected alignment or sentence parsing, punctuation was normalized where possible and documented. Some texts, e.g., *A volta* or *Cada raça tem um Calino*, were especially affected by long or complex sentences, requiring sentence splitting, deletions, or insertions of artificial breaks to enable further processing and in extension syntactic comparison. The following list illustrates key issues and interventions.

- Normalization of punctuation to align segmentation between Stanza language models: Semicolons were replaced with commas where appropriate to maintain consistent sentence segmentation across all versions.
- Dialogue marker removal: Dialogue markers such as hyphens or quotation marks disrupted segmentation and were uniformly removed.
- Sentence filtering: Sentences that could not be aligned due to parser errors, unjustified segmentation by Stanza, or translation-induced length mismatches were removed. Unjustified segmentation includes cases where the Portuguese language model split sentences incorrectly, for instance when a capitalized named entity was misinterpreted as the beginning of a new sentence (see 7.2.1 for a detailed explanation).
- Additional formatting: Artificial line breaks were inserted after colons or ellipsis to aid segmentation.

Based on the adjusted and aligned sentence pairs, stored in Tab-Separated Values (TSV) file format, various corpus statistics were calculated. Metrics such as BLEU, TER, chrF, and COMET used the HT or PEMT as reference, while XCOMET relied solely on the ST. Three TSV files per text were created. One TSV file comprises MT aligned to HT/PEMT, thus, resulting in a monolingual German-to-German parallel data file. The second file contains the bilingual Portuguese-to-German

parallel data, ST to HT/PEMT, which might help in future research on translation of literature. The third TSV file contains a bilingual version as well, ST to MT.

All sentence-aligned TSV files were parsed using the Stanza NLP pipeline, generating linguistic annotations for each sentence. For both Portuguese and German texts, this included tokenization, lemmatization, POS tagging, dependency relation parsing, and constituency parsing. Thus, each sentence was enriched with syntactic and morphological information, which serves as the basis for further alignment-based syntactic evaluation.

7.2.1. Word Alignment

Table 9 shows an example sentence from the dataset with the alignment results obtained using the three different alignment strategies by SimAlign. The sentences are translations of the very first sentence of Machado de Assis' text *A Cartomante*. This setup allows us to examine how the machine-generated sentence is aligned with its human counterpart and to investigate structural and syntactic differences between MT and HT. This sentence pair is particularly interesting because it contains multiple clauses for which there is a notable syntactic divergence between the MT and HT.

Table 9.: Example of a sentence alignment of a MT-HT pair with different alignment strategies

Type	**Content**
MT	*HAMLET bemerkt gegenüber Horatio , dass es mehr Dinge zwischen Himmel und Erde gibt , als unsere Philosophie sich träumen lässt .*
HT	*Es gibt mehr Dinge zwischen Himmel und Erden , als unsere Schulweisheit sich träumen lässt , sagt Hamlet zu Horatio .*
mwmf alignments	0-11 0-17 1-7 1-16 2-18 3-19 4-15 6-0 7-2 8-3 9-4 10-5 11-6 12-7 13-1 14-8 15-9 16-10 17-11 18-12 19-13 20-14 21-20
inter alignments	1-16 2-18 3-19 6-0 7-2 8-3 9-4 10-5 11-6 12-7 13-1 14-8 15-9 16-10 17-11 18-12 19-13 20-14 21-20
itermax alignments	0-17 1-11 1-16 2-18 3-19 6-0 7-2 8-3 9-4 10-5 11-6 12-7 13-1 14-8 15-9 16-10 17-11 18-12 19-13 20-14 21-20

To illustrate the alignment behavior of different strategies, Fig. 13, Fig.12, and Fig.14 show the same MT-HT sentence pair aligned using the Intersection, MWMF, and Itermax strategies, respectively. These visualizations make it easier to observe

how the different strategies handle complex sentence structure and long-distance dependencies. The MWMF strategy produces the most alignments, capturing both core and peripheral correspondences, while the Itermax strategy yields slightly fewer but still extensive alignments. The Intersection strategy, as expected, results in the fewest alignments, reflecting its conservative approach of retaining only those alignments agreed upon by multiple internal heuristics. This illustrates the trade-offs between coverage and precision. While MWMF maximizes the number of aligned words, which leads to a higher coverage, Intersection emphasizes precision at the cost of leaving some words unaligned. Itermax strikes a balance between these two extremes.

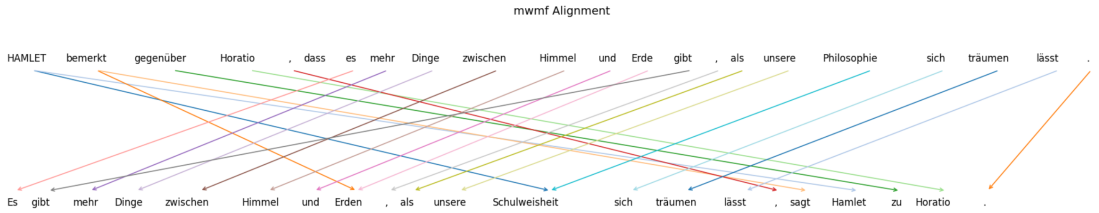

Figure 12.: MWMF alignment strategy applied to a sample sentence

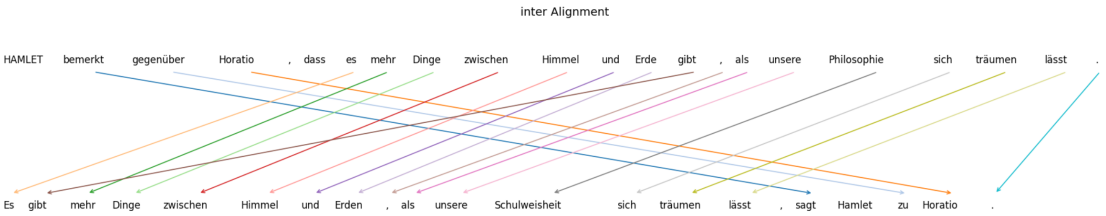

Figure 13.: Intersection alignment strategy applied to a sample sentence

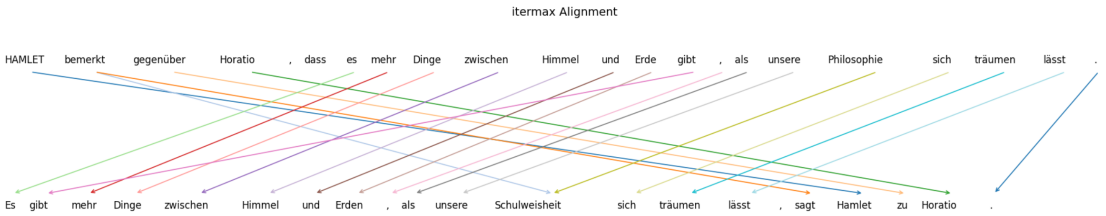

Figure 14.: Itermax alignment strategy applied to a sample sentence

Although each strategy has its strengths and weaknesses, the intersection method was chosen for all subsequent evaluations. As illustrated in Fig. 13, this strategy produces fewer alignments in general, which can lead to lower coverage. However, it is also more conservative, resulting in fewer wrong links. This makes it particularly suitable for syntactic evaluation tasks where the precision of alignments is more critical than sheer quantity.

During preparation of the dataset for syntactic evaluation, a small number of sentence pairs had to be excluded due to malformed dependency trees. Especially ASTrED relies on well-formed trees with a single root to compute tree edit distances accurately. However, some sentence pairs contained multiple roots, rendering them incompatible with ASTrED. This often happened due to parsing errors or irregular sentence structures. Table 10 lists all affected files, indicating how many sentence pairs were skipped and how many remained usable. For example, in the parallel data file *Qualquer serve* MT to HT, 1 of 30 pairs of sentences was removed due to the mapping of two roots, also called heads, in the MT. This particular sentence assigns to two verbs the dependency label *root*, which violates the requirement of a single root. Such issues were identified through head index analysis and filtered out prior to evaluation.

Table 10.: Files with malformed dependency trees due to multiple or missing roots

File	**Total**	**Skipped**	**Usable**
a-cartomante_mt_ht	199	1	198
a-cartomante_st_ht	199	1	198
a-cartomante_st_mt	199	2	197
cada-raça-tem-um-calino_mt_ht	83	1	82
cada-raça-tem-um-calino_st_ht	83	2	81
cada-raça-tem-um-calino_st_mt	83	1	82
linhas-de-tiro_mt_ht	24	1	23
linhas-de-tiro_st_ht	24	1	23
qualquer-serve_mt_ht	30	1	29
qualquer-serve_st_mt	30	1	29
uma-lembrança_mt_ht	21	2	19
uma-lembrança_st_ht	21	2	19
uma-lembrança_st_mt	21	2	19

Each sentence is represented as a list of tokens and a corresponding list of syntactic heads, following the UD convention. The head of a word is the word that it syntactically depends on, often its grammatical parent. These heads are encoded as a list of integers, where each number represents the position of the parent word in the sentence. The indexing starts at 1, and the special value 0 is reserved for

the root word, which does not have a parent. In this structure, a value of 3 means that a word depends on the third word in the sentence, and a value of 0 indicates that the word is a root. If a sentence has more than one word with head index 0, it means that multiple roots were detected, rendering the structure invalid for ASTrED. The sentence in Example (1) shows this structure with the identified roots in bold. In the source tree, both *dachte* at index 3 and *blieb* at index 10 are marked as roots, making the structure invalid for ASTrED evaluation. In contrast, the target sentence correctly has only one root, *dachte* at index 3.

(1) (a) **Source:**

Der	Staatsanwalt	**dachte**	einige	Minuten	nach	;	der
2	3	**0**	5	3	3	3	2

Baron	**blieb**	stehen	und	wartete	auf	die	Antwort
3	**0**	3	6	4	9	9	6

des	jungen	Mannes	,	bis	dieser	sagte	:
12	12	9	16	16	16	6	3

(b) **Target:**

Der	Staatsanwalt	**dachte**	einige	Minuten	lang	nach	,
2	3	**0**	5	3	3	3	11

der	Baron	wartete	mit	Spannung	auf	die	Antwort
10	11	3	13	11	16	16	11

des	jungen	Mannes	,	bis	dieser	sagte	:
19	19	16	23	23	23	11	3

7.2.2. Syntactic Diversity Metrics

To capture syntactic variation across translation modes, several complementary metrics are applied on the basis of the previously established word alignments. The first measure is the count of word label changes, which compared the dependency labels of aligned tokens. For each aligned word pair produced by SimAlign, the syntactic function of the source tokens is compared to that of the target token. A word label change is counted whenever the two labels differ, for instance, if a token labeled as a subject in the source sentences is aligned to a token labeled as an object in the target sentence.

The second set of measures focuses on crossings, which quantify reordering by examining whether aligned word pairs intersect when mapped across the two sentences. Three variants are calculated. The moist basic form, word crosses, counts intersecting alignment links directly, providing a raw indication of reordering.

Sequence crossings extend this principle by grouping consecutive words into units, which allows the metric to capture larger reordering phenomena beyond single-word movements. Building on this, SACr introduces a linguistic refinement in which groups are only considered valid if they form subtrees in the dependency structure, e.g., all words within a group share direct parent-child relations. If a group does not form a valid subtree, it is split into smaller ones. By grounding reordering in syntactic structure, SACr provides a more meaningful interpretation, since surface shifts in word order do not always correspond to genuine syntactic restructuring. In each case, the number of crosses is normaized by the total number of alignments, resulting in a score where higher values indicate greater reordering. Finally, ASTrED is applied to assess structural divergence at the level of entire dependency trees. ASTrED calculates the minimal number of edit operations required to transform one tree into the other, constrained by the word alignments provided by SimAlign. These operations include insertions, deletions, and substitutions of nodes or labels, and yield a measure of global syntractic divergence between two sentences. Sentences containing malformed trees were excluded from this step, since ASTrED requires well-formed trees. The reason for the malformation of some trees was the presence of multiple roots, most often caused by parsing errors that assigned the dependency label *root* to more than one token. In such cases, the structure violates the UD requirement of a single root and is incompatible with ASTrED, which assumes tree-shaped structures. These instances were systematically identified and filtered prior to evaluation, ensuring that all ASTrED computations were performed only on valid dependency trees.

7.3. Manual Inspection

To complement the quantitative analysis, a small number of aligned sentence pairs are selected for close manual inspection. The goal is to validate the reliability of the automatically computed metrics by checking whether observed numerical trends correspond to actual syntactic differences in the text. It also provides more nuanced insight into how MT, PEMT, and HT handle syntactic restructuring in literary contexts. The inspected examples were chosen primarily from outlier texts that display exceptionally high or low values in one or more syntactic metrics to explore the causes of these deviations. Focusing on these outliers helps identify specific structural features, such as extensive reordering and complex subordination, that account for the deviations observed in the quantitative results.

8. Results

To establish a baseline for translation quality, a set of standard evaluation metrics was calculated across all 24 texts. Fig. 15 presents the scores for BLEU, TER, and chrF, with the exact numerical values in Table 11. For BLEU and chrF, a higher score indicates a closer n-gram or character-level overlap with the reference translation and thus a higher assumed translation quality. In contrast, TER is interpreted inversely, where lower values indicate fewer required edits and therefore better translation quality. These three measures illustrate different aspects of surface-level correspondence between ST, MT and PEMT for *O Pavilhão e a Pinel* and between ST, MT and HT for all other texts.

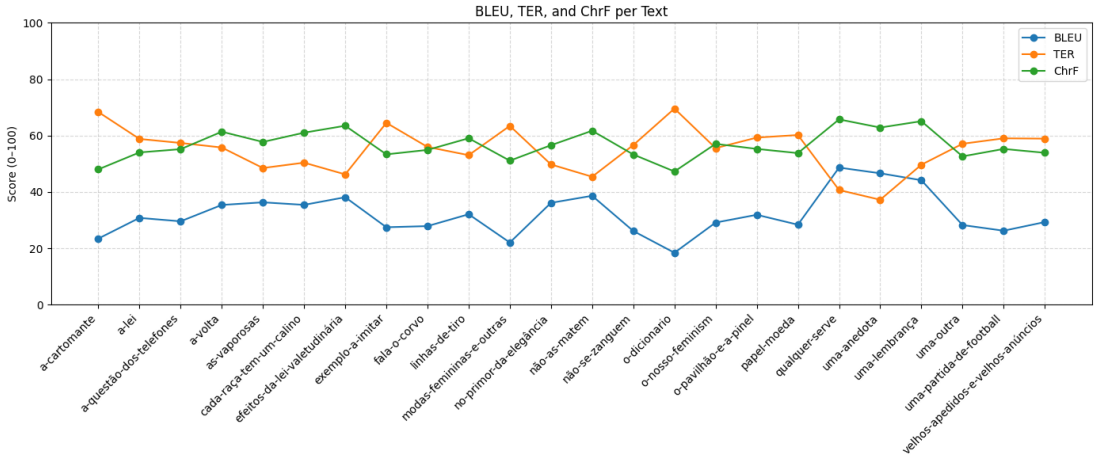

Figure 15.: Visualization of the metrics BLEU, TER, and chrF

BLEU scores range from a minimum of 18.43 for *O dicionario* to a maximum of 48.63 for *Qualquer serve*. TER values show an inverse trend, with the lowest editing effort of 37.22 for *Uma anedota* and the highest of 69.56 for *O dicionario*. ChrF generally follows the same tendencies as BLEU, but on a different scale, ranging from 47.25 for *O dicionario* to 65.73 for *Qualquer serve*. Overall, the three metrics reveal that certain texts, e.g, *Qualquer serve, Uma andedota, Uma lebrança*, achieve relatively high lexical overlap and low editing effort, whereas others, e.g., *O dicionario, Modas femininas e outras, A cartomante*, consistently perform worse across all surface-level metrics. Interestingly, as also visible in Fig. 15, the three

Table 11.: Translation quality metrics BLEU, TER, and chrF across all 24 texts

Text	**BLEU**	**TER**	**chrF**
A cartomante	23.44	68.36	48.00
A lei	30.79	58.80	54.01
A questão dos telefones	29.58	57.37	55.17
A volta	35.38	55.76	61.36
As vaporosas	36.31	48.48	57.73
Cada raça tem um calino	35.41	50.38	61.02
Efeitos da lei valetudinária	38.12	46.22	63.47
Exemplo a imitar	27.47	64.51	53.32
Fala o corvo	27.88	55.95	54.91
Linhas de tiro	32.08	53.04	59.04
Modas femininas e outras	22.05	63.43	51.08
No primor da elegância	36.13	49.74	56.57
Não as matem	38.63	45.39	61.70
Não se zanguem	26.14	56.62	53.28
O dicionário	18.43	69.56	47.25
O nosso feminism	29.13	55.54	57.05
O pavilhão e a Pinel	31.89	59.27	55.26
Papel-moeda	28.38	60.19	53.77
Qualquer serve	48.63	40.68	65.73
Uma anedota	46.61	37.22	62.83
Uma lembrança	44.15	49.65	65.09
Uma outra	28.25	57.08	52.61
Uma partida de football	26.24	59.01	55.25
Velhos apedidos e velhos anúncios	29.31	58.90	53.88
Average	**32.10**	**55.05**	**56.64**

metrics show a clear correlation. The BLEU and chrF scores follow nearly identical trends across the dataset, while the TER curve moves in the opposite direction. This inverse relationship confirms that higher BLEU and chrF scores correspond to lower TER values, reflecting that translations with greater lexical overlap with the reference require fewer edits overall.

In addition to these surface-oriented metrics, Fig. 16 reports the results for COMET and XCOMET, which are based on neural regression models trained to approximate human judgment of translation quality. The exact numerical values can be seen in Table 12. While COMET is optimized for the evaluation of machine output against human reference translations, XCOMET extends the approach to allow comparisons without requiring references. Fig. 16 therefore shows one

COMET score and two XCOMET scores, for MT and PEMT (*O Pavilhão e a Pinel*), and MT and HT respectively, for each text.

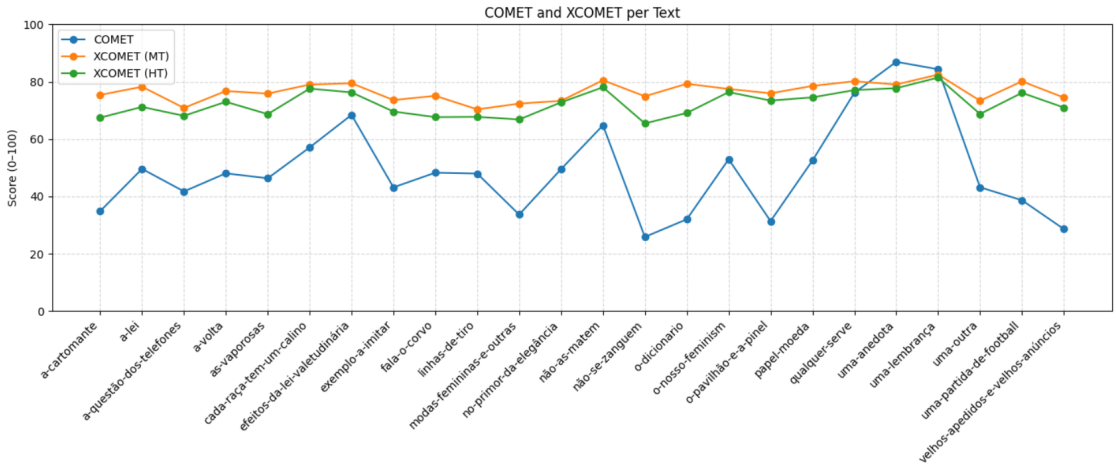

Figure 16.: Visualization of the metrics COMET and XCOMET

The COMET scores have a wider spread, ranging from 25.87 for *Não se zanguem* to 86.94 for *Uma anedota*. Notably, the highest COMET scores align with the texts that also performed strongly in BLEU/chrF. XCOMET, which evaluates without references, produces overall higher scores. MT scores range between 70.36 for *Linhas de tiro* and 82.48 for *Uma lembrança*, while HT scores are slightly lower on average, ranging from 65.45 for *Não se zanguem* to 81.44 for *Uma lembrança*. Interestingly, MT scores often exceed HT scores, suggesting that the system output in all cases used in this study is closer to the training distribution of the evaluation model than the HT.

Table 13 provides an overview of the average values obtained for the syntactic evaluation metrics across the corpus. The results confirm a consistent pattern across all measures: MT-HT pairs yield the lowest number of changes, while ST-HT pairs show the highest, with ST-MT values falling in between but still closer to the latter than the former. This reflects the different nature of the comparisons with two German texts in the MT-HT dataset, leading to fewer differences, whereas ST-based comparisons necessarily involve cross-linguistic restructuring.

Table 12.: COMET and XCOMET scores for each text

Text	**COMET**	**XCOMET MT**	**XCOMET HT**
A cartomante	34.86	75.39	67.47
A lei	49.60	78.25	71.20
A questão dos telefones	41.74	70.86	68.10
A volta	48.05	76.75	72.99
As vaporosas	46.32	75.86	68.69
Cada raça tem um calino	57.08	79.01	77.60
Efeitos da lei valetudinária	68.44	79.47	76.28
Exemplo a imitar	43.16	73.60	69.59
Fala o corvo	48.28	75.08	67.65
Linhas de tiro	47.96	70.36	67.75
Modas femininas e outras	33.66	72.36	66.83
No primor da elegância	49.51	73.34	72.80
Não as matem	64.80	80.52	78.06
Não se zanguem	25.87	74.92	65.45
O dicionário	32.02	79.28	69.09
O nosso feminism	52.95	77.47	76.37
O pavilhão e a Pinel	31.39	75.95	73.42
Papel-moeda	52.51	78.53	74.54
Qualquer serve	76.13	80.16	77.08
Uma anedota	86.94	79.03	77.74
Uma lembrança	84.34	82.48	81.44
Uma outra	43.18	73.37	68.75
Uma partida de football	38.66	80.14	76.12
Velhos apedidos e velhos anúncios	28.61	74.47	71.00
Average	**49.42**	**76.53**	**72.33**

Table 13.: Average syntactic evaluation metrics across the corpus

Metric	**MT–HT**	**ST–MT**	**ST–HT**
Dependency Label Changes	2.49	4.45	4.57
POS Tag Changes	1.24	2.70	2.77
Word Cross	4.20	5.94	6.12
Sequence Crossings	1.54	3.26	3.60
SACr	2.17	3.89	4.31
ASTrED	7.81	11.35	11.79

8.1. Word Label Changes and Crosses

In addition to general purpose evaluation metrics such as BLEU or COMET, this study applies syntactically informed measures to capture structural differences between MT, HT, and PEMT outputs. Figures 17, 18, 19, 20 and 21 present the results for dependency label changes, POS-tag changes, and three variants of crossing measures, word-level, sequence-level, and syntactically aware sequence-level. Together, these metrics provide a more fine-grained picture of translation diversity at the syntactic level.

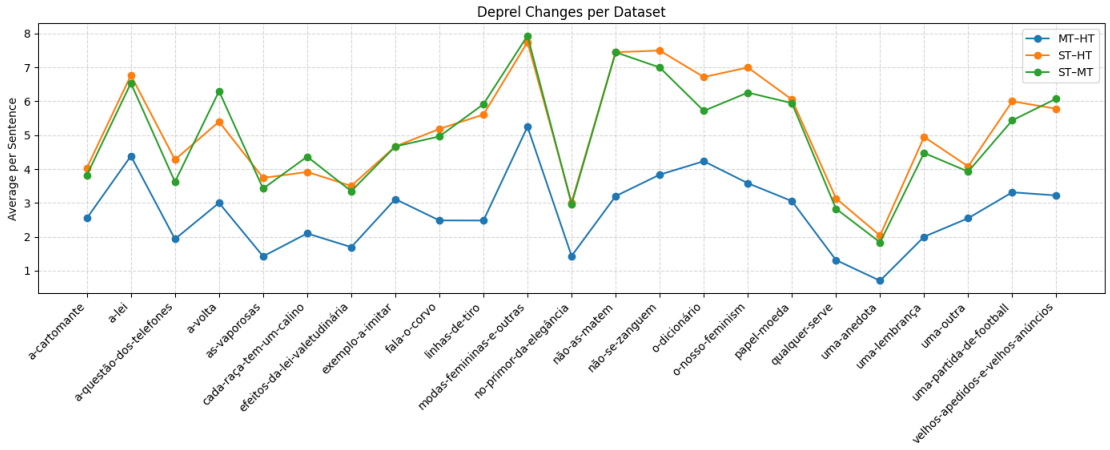

Figure 17.: Visualization of dependency changes across the dataset

Figure 17 shows the distribution of the changes in the dependency labels across the datasets. MT-HT pairs exhibit the lowest range of variation from 0.70 to 5.52, followed by ST-HT from 2.03 to 7.75 and ST-MT from 1.83 to 7.94. The lowest bounds in all cases are found in *Uma anedota*, while the upper bounds occur in *Modas femininas e outras*. Overall, MT-HT pairs consistently show the lowest number of dependency changes, which can be attributed to the fact that both sides of the comparison are in the same language. In contrast, the ST-HT and ST-MT pairs exhibit considerably more variation in relation to dependency labels. The two are generally close to each other, with ST-HT tending to show more changes than ST-MT. Nevertheless, five texts show the opposite pattern, where ST-MT displays more changes than ST-HT, and in two texts the values are identical for ST-MT and ST-HT. This suggests that syntactic restructuring is generally more pronounced in HTs, while MTs tend to stay closer to the source. However, the proximity of the ST-HT and ST-MT averages, as well as individual cases where ST-MT exceeds ST-HT, suggests that MT is not uniformly conservative. In certain texts, MT introduces a degree of syntactic divergence comparable to or even greater than that of HT. For the single PEMT sample included in the dataset, the MT-PEMT

comparison yields 3.57 dependency label changes, which is higher than the MT-HT average of 2.49. Similarly, ST-PEMT shows 6.62 changes, compared to an average of 4.57 for ST-HT, while the corresponding ST-MT value of 5.73 is also above the overall ST-MT average of 4.45.

In order to contextualize the dependency label change patterns, the relative frequencies of the different labels were examined (see Appendix A.1 for exact values). Relative rather than absolute frequencies are reported, as the number of sentences differs across modalities. The most frequent labels in all modalities are punctuation, determiners, case markers or prepositions, nominal subjects, and adverbial modifiers, which together account for roughly half of all dependencies. At a finer level, some differences between modalities emerge nonetheless. ST generally exhibits higher frequencies of determiners, case markers, nominal modifiers, and markers, whereas the translations tend to increase auxiliaries, clausal modifiers, and certain relational elements such as conjunctions and oblique arguments. PEMT often amplifies these shifts, showing the highest frequencies for conjunctions and obliques, while roots appear slightly less frequent. Overall, these patterns suggest that translation preserves the core sentence structure but certain dependency types are systematically adjusted across modalities, especially in PEMT.

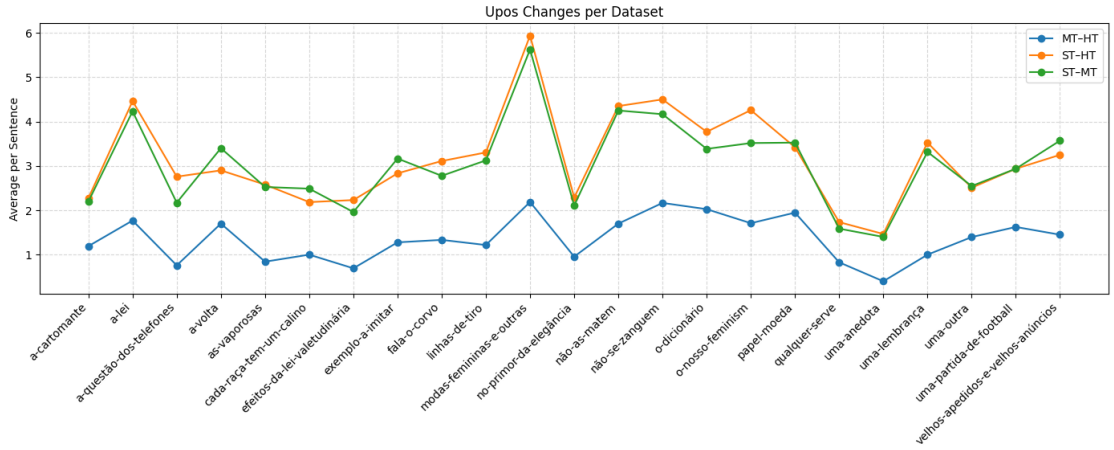

Figure 18.: Visualization of POS-tag changes across the dataset

Figure 18 displays the distribution of POS-tag changes in the datasets. For MT-HT, the values range from 0.40 to 2.19. ST-HT pairs show a broader span from 1.47 to 5.94, while ST-MT pairs vary between 1.40 and 5.62. The lower and upper bounds in all cases are represented by the same two texts as before, with *Uma anedota* at the lower end and *Modas femininas e outras* at the upper end. Overall, MT-HT comparisons produce the lowest values, reflecting fewer changes in the POS-tags, which can be attributed to the fact that the same language is used

on both sides, leading to less syntactical differences. ST-HT and ST-MT values are considerably higher and largely overlap, with ST-HT tending to exceed ST-MT. In six cases, however, ST-MT exhibits more changes in POS-tags compared to ST-HT, and in one case the values are identical. For the PEMT sample, MT-PEMT records 1.40 POS-tag changes, compared to an average of 1.24 for MT-HT. ST-PEMT reaches 3.58, which is higher than the ST-HT average of 2.77.

The distribution of POS-tags reveals a high degree of structural consistency across all modalities, with nouns, punctuation marks, determiners, verbs, pronouns, and adpositions forming the dominant categories (see Appendix A.1 for numerical values). These six POS classes together account for more than 75% of all tokens in each modality. Relative differences between the ST and its translations show subtle shifts in syntactic profile. The ST tends to have fewer pronouns and adjectives but slightly more verbs, adpositions, and subordinating conjunctions, while the translations generally increase pronouns and adjectives, with PEMT showing highest frequencies. Auxiliaries are lowest in the ST, equal in HT and PEMT, and slightly higher in MT, and particles, which are absent in the ST appear consistently in the translations. Determiners decrease from ST to MT to HT down to PEMT. Overall, these patterns show that while the translations retain the core syntactic composition of the ST, there are small shifts in certain categories, with PEMT generally exhibiting the strongest deviations from the ST.

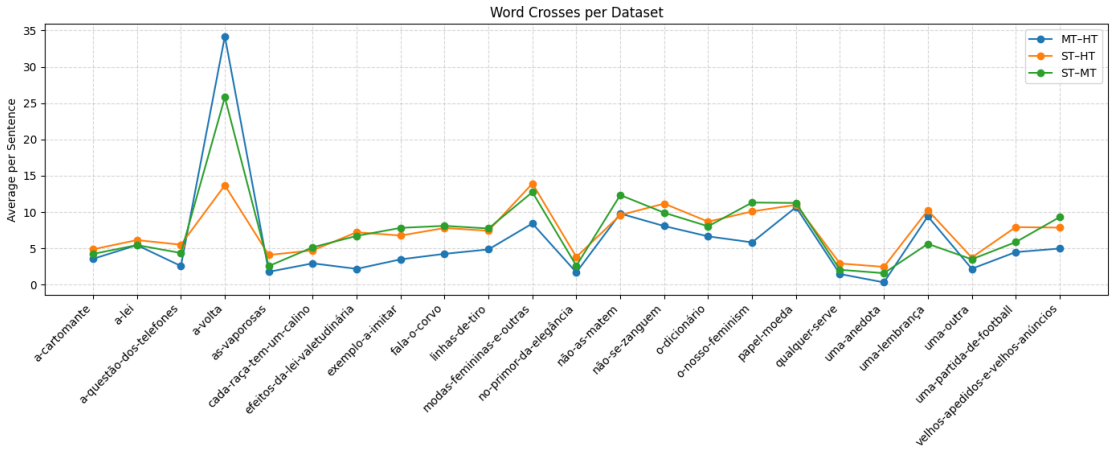

Figure 19.: Visualization of word crosses across the dataset

Figure 19 illustrates the number of crossing alignment links at the word level. MT-HT values range widely, from as low as 0.33 to as high as 34.20, reflecting extreme variation in how closely aligned MT is to human translation word order. ST-HT comparisons show a more narrow range between 2.47 and 13.94, while ST-MT covers a similar but sometimes more extreme range, from 1.60 to 25.80.

On average, MT-HT shows the lowest number of crosses with 4.20, while ST-based comparisons show more frequent crossing alignment links with averages of 5.94 for ST-MT and 6.12 for ST-HT. In eight of the 23 cases, the ST-MT values appear higher than the ST-HT values. Additionally, one case seems to have identical values for ST-MT and ST-HT and another case presents MT-HT as the highest value. For the PEMT sample, MT-PEMT records 6.07 crosses, which is above the MT-HT average, while ST-PEMT reaches 10.51, substantially higher than both ST-HT and ST-MT. This pattern suggests that PE introduces additional word-level reordering beyond those found in either MT or HT, indicating that editors intervene strongly in word order.

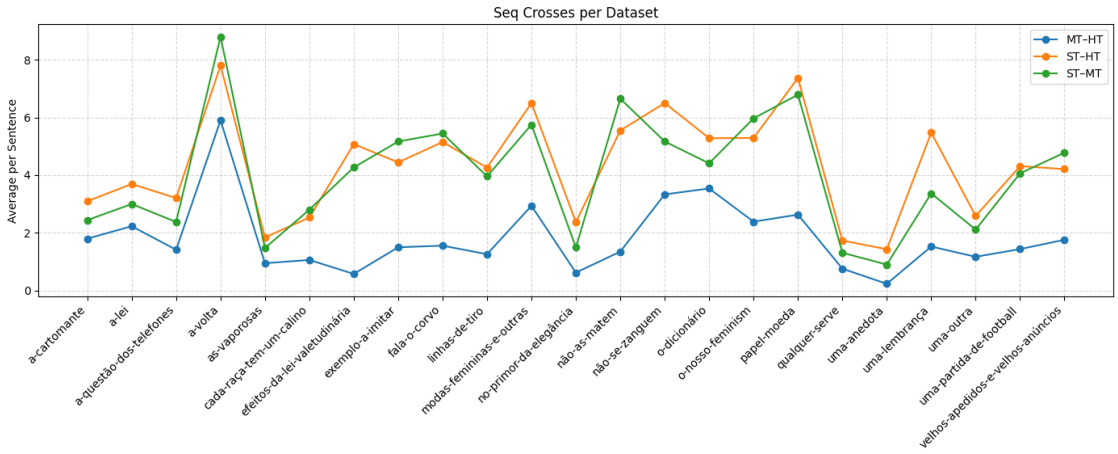

Figure 20.: Visualization of sequence crosses across the dataset

Figure 20 tracks sequence-level crossing links, where blocks of words rather than single tokens change order. MT-HT values range from 0.23 to 5.90. For ST-HT, the range is higher, from 1.43 to 7.80. ST-MT shows a similar but slightly wider distribution, from 0.90 to 8.80. In all three cases *Uma anedota* records the lowest number of sequence crosses, while *A volta* produces the highest. This pattern largely mirrors the word-level crossings, with MT-HT values remaining the lowest at 1.54 on average, whereas ST-HT at 3.60 and ST-MT at 3.26 track more closely with each other, with HT generally more restructuring. In six cases, however, ST-MT appears to have more sequence crossing links than ST-HT, suggesting that MT occasionally produces more reordering than HT. The PEMT sample reinforces this picture. MT-PEMT yields 2.59 crosses, which is higher than the MT-HT average, while ST-PEMT reaches 5.94, exceeding both ST-HT and ST-MT. This indicates that post-editors introduce substantial reordering of word sequences, going beyond both MT's tendency to remain closer to the source and HT's restructuring strategies.

Figure 21 presents syntactically aware crosses, which combine reordering with the dependency structure. MT-HT values range between 0.27 and 13.20, with an average of 2.17. ST-HT pairs span from 1.80 to 8.90, the average being 4.31, while the ST-MT values fall between 1.00 and 10.70 and an average of 3.89. On average, ST-HT exceeds ST-MT, confirming that human translators engage in deeper syntactic restructuring than MT, though certain MT-HT comparisons, such as *A volta*, reveal high divergence. *Uma anedota* is again the case with the lowest crossing links. The same exceptions as in Figure 20 can also be seen here. For the PEMT sample, MT-PEMT records 3.60, higher than the MT-HT average, while ST-PEMT reaches 7.17, exceeding both ST-HT and ST-MT. This suggests that post-editors introduce more syntactically aware crossings than either MT or HT, reflecting their tendency to correct surface-level issues while retaining MT's underlying syntactic framework.

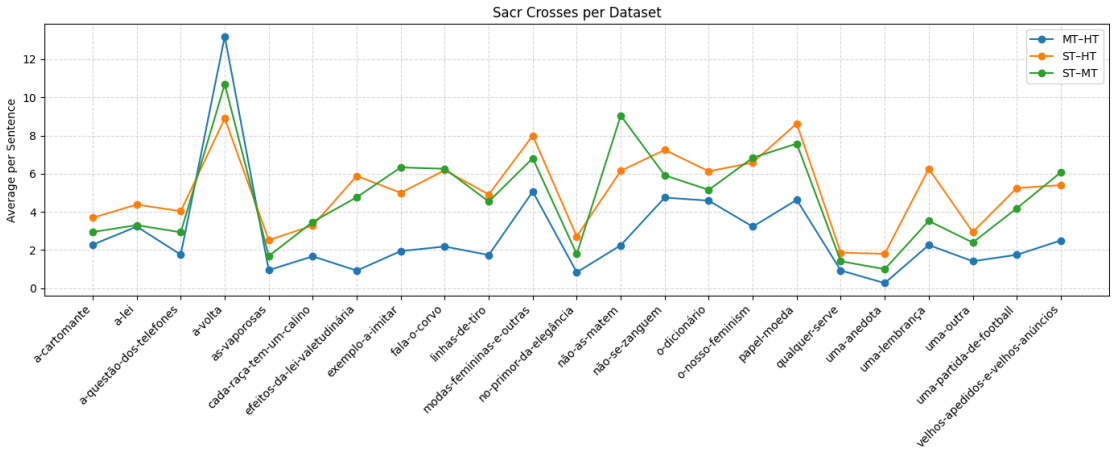

Figure 21.: Visualization of syntactically aware crosses across the dataset

8.2. ASTrED

Figure 22 displays the ASTrED scores, which measure overall syntactic divergence by integrating tree structure and word alignment. MT-HT pairs show the lowest values, ranging from 2.03 to 12.88, with an average of 7.81. ST-HT comparisons cover a higher span, from 4.70 to 17.77, averaging 11.79. ST-MT pairs fall within a similar range, from 4.00 to 18.69, with an average of 11.35. Again for all three sets *Uma anedota* is at the bottom. *Modas femininas e outras* has the highest values for MT-HT and ST-MT, while for ST-HT *A lei* presents the upper end. Eight cases of ST-MT present higher numbers than ST-HT. This distribution confirms the consistent pattern observed across the other metrics. MT-HT scores remain

low, while ST-based comparisons yield considerably higher divergence, with ST-HT tending to exceed ST-MT. As for the PEMT sample, MT-PEMT reports 11.18, substantially higher than the MT-HT average. ST-PEMT reaches 15.75, compared to 13.75 for ST-MT of that text and 11.79 for the average ST-MT.

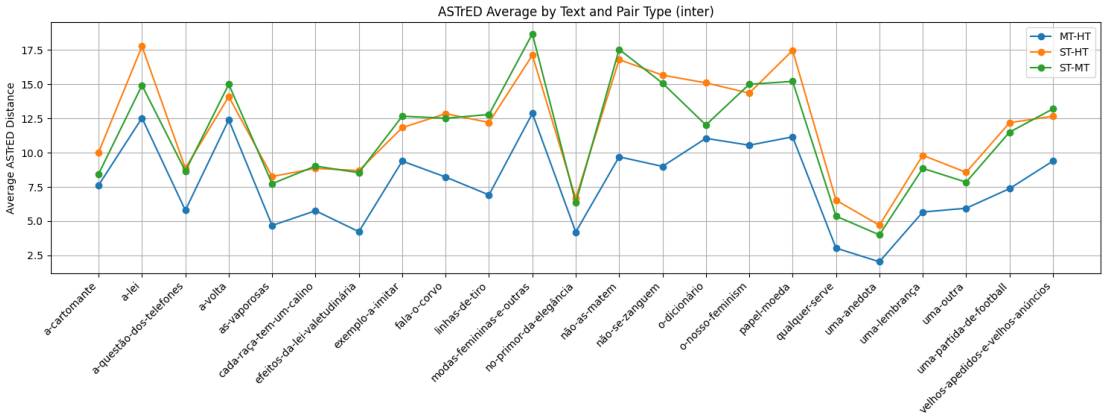

Figure 22.: Visualization of aligned syntactic tree edit distance across texts

8.3. Manual Inspection

To illustrate how the syntactic metrics correspond to actual translation phenomena, a selection of sentences was examined. Figure 23, Figure 24, and Figure 25 show visualizations of word alignments for one sentence from *Uma Anedota* across the MT-HT, ST-HT, and ST-MT datasets. This sentence was selected because the text consistently showed low syntactic divergence, providing a clear example of subtle structural differences.

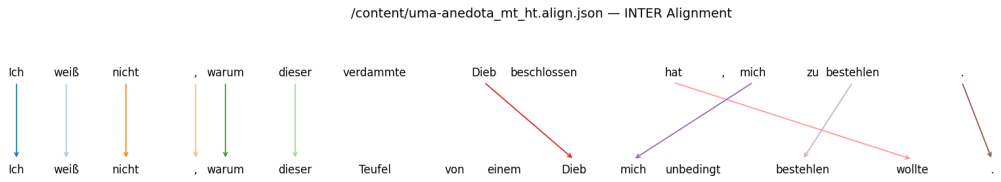

Figure 23.: Alignment visualization of the first sentence in *Uma anedota*, MT-HT

Table 14 summarizes the syntactic metrics for this sentence across the three dataset pairs. The overall alignment coverage is stable, 11 links in each case, but the number of WLC and POS-tag changes varies. MT-HT shows only two label

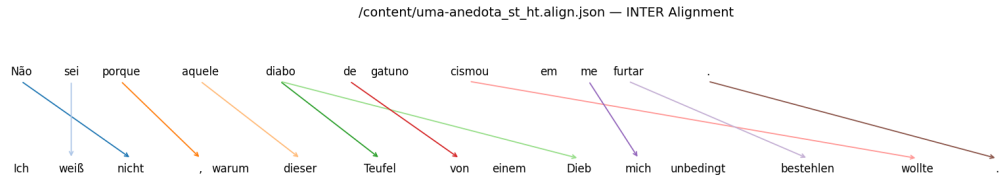

Figure 24.: Alignment visualization of the first sentence in *Uma anedota*, ST-HT

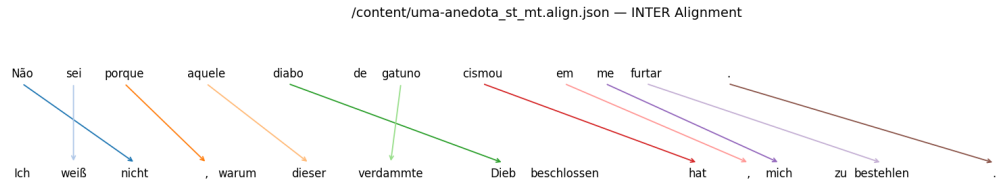

Figure 25.: Alignment visualization of the first sentence in *Uma anedota*, ST-MT

changes and no POS-tag changes, while ST-HT increases to four label changes and three POS-tag changes, and ST-MT displays the highest divergence with five changes in both categories.

Table 14.: Syntactic metrics for sentence 1 from *Uma Anedota*. Total alignments, WLC changes, and POS-tag changes are shown for each dataset pair

Dataset Pair	Total Alignments	WLC	POS-tag Changes
MT–HT	11	2	0
ST–HT	11	4	3
ST–MT	11	5	5

A closer inspection of the MT-HT pair illustrates how the metrics reflect specific syntactic choices, see Table 15. Out of 11 alignments, two involved dependency label changes: the word *Dieb* ('thief') is labeled as a subject (*nsubj*) in the MT version but as an oblique (*obl*) in the HT version, while the verb *bestehlen* ('to rob') is annotated as an open clausal complement (*xcomp*) in MT but as a clausal complement (*ccomp*) in HT.

The remaining alignments show strong syntactic stability. For instance, the root verb *weiß* ('know'), the negation *nicht* ('not'), and the object *mich* ('me') align perfectly with identical dependency labels in both versions. At the POS level, no mismatches were found as all aligned tokens shared identical UPOS categories, e.g., *Ich* ('I') as *PRON*, *weiß* ('know') as *VERB*.

Table 15.: Examples of dependency relation (deprel) and POS-tag matches/mismatches for sentence 1 (MT-HT, *Uma Anedota*)

MT (Index, label)	**HT (Index, label)**	**Relation**
Dependency label mismatches		
Dieb (7, nsubj)	Dieb (9, obl)	nsubj vs. obl
bestehlen (13, xcomp)	bestehlen (12, ccomp)	xcomp vs. ccomp
Dependency label matches		
Ich (0, nsubj)	Ich (0, nsubj)	nsubj
weiß (1, root)	weiß (1, root)	root
nicht (2, advmod)	nicht (2, advmod)	advmod
, (3, punct)	, (3, punct)	punct
warum (4, mark)	warum (4, mark)	mark
dieser (5, det)	dieser (5, det)	det
hat (9, aux)	wollte (13, aux)	aux
mich (11, obj)	mich (10, obj)	obj
. (14, punct)	. (14, punct)	punct
POS matches		
Ich (0, PRON)	Ich (0, PRON)	PRON
weiß (1, VERB)	weiß (1, VERB)	VERB
nicht (2, PART)	nicht (2, PART)	PART
, (3, PUNCT)	, (3, PUNCT)	PUNCT
warum (4, ADV)	warum (4, ADV)	ADV
dieser (5, DET)	dieser (5, DET)	DET
Dieb (7, NOUN)	Dieb (9, NOUN)	NOUN
hat (9, AUX)	wollte (13, AUX)	AUX
mich (11, PRON)	mich (10, PRON)	PRON
bestehlen (13, VERB)	bestehlen (12, VERB)	VERB
. (14, PUNCT)	. (14, PUNCT)	PUNCT

In contrast, *A volta* shows the highest values across the syntactic crossing metrics, which reflects substantial structural reorganization between the ST and the translations, especially where MT is involved. While most sentence pairs show relatively consistent structures, a few sentences are made up of long and clause-dense constructions which tend to trigger reordering during translation. Interestingly enough, this trend runs counter to the general expectation that MT would remain structurally closer to the ST than HT. In these few complex sentences, MT diverges more from the source syntax, while HT maintains a closer alignment. In addition the values are amplified by the fact that this text consists of only ten sentences overall, of which three display particularly complex constructions, thereby exerting a disproportionate influence on the average syntactic scores.

In *Modaas femininas e outras*, which comprises 16 sentences, the syntactic analysis reveals high values across all metrics. The text shows the highest numbers of dependency label and POS-tag changes among all analyzed texts, alongside the highest ATReD scores for MT-HT and ST-MT comparisons. The ST-HT pair also displays the highest number of word cross links, indicating substantial reordering between source and HT. The text itself, a humorous and rhethorically elaborate reflection on fashion, gender, and morality, contains long, intricately structured sentences that pose challenges to both translation types. The frequent label and

POS-tag shifts suggest that both, the machine and the human translator, often resorted to structural and grammatical reformulation, reflecting the need to adapt the syntax and lexical variety of the source. These findings suggest that both MT and HT diverge considerably from the ST, but in slightly different ways. MT shows a greater overall syntactic distance from the source according to ASTrED scores, whereas HT introduces extensive word reordering.

The PEMT *O Pavilhão e a Pinel*, which consists of 88 sentences, exhibits consistently above-average values across all syntactic metrics, without reaching the extreme highs or lows observed in other texts. As this is the only text representing the PEMT modality, and given that it contains a number of long and syntactically dense sentences, its slightly elevated values are unsurprising. The source material itself is dense, introspective, and linguistically elaborate, alternating between fragmented reflections and long, complex sentences. The text's behavior reflects PE practice in that it largely preserves the structural organization of the machine output while implementing targeted revisions to improve grammar and fluency. Most changes concern local adjustments rather than full reorganization, such as fine-tuning word order within clauses, modifying dependency relations, or replacing literal lexical choices with more natural alternatives. As a result, the PEMT achieves smoother and more coherent surface structure than the raw MT output while still retaining elements of its syntactic framework, which explains its moderate but consistent divergence across all metrics.

Taken together, this section complements the quantitative results by illustrating how syntactic metrics correspond to observable linguistic phenomena. Across texts, higher metric values tend to coincide with complex source structures, long or multi-clause sentences, or even stylistically elaborate passages that trigger syntactic reorganization. Conversely, lower scores often reflect simpler syntactic patterns or close structural correspondence between texts.

9. Discussion

This thesis aimed to investigate the syntactic diversity between the HT, MT, and PEMT of literary texts, with a particular focus on whether PEMT aligns more closely with MT or HT. Previous research on technical and non-literary domains has shown that MT output typically remains structurally closer to the ST, while the HT demonstrates greater syntactic reorganization (Toral, 2019; Shaitarova et al., 2023). The present study extends this line of inquiry to the literary domain. The results consistently showed that MT-HT comparisons involved the least divergence, which was expected since both sides of the comparison are in German and no cross-lingual transfers is involved. Comparisons involving the ST revealed considerably more restructuring, with HTs generally displaying higher degrees of syntactic variation than MTs. This confirms the tendency of MT to preserve source-language structures, while HTs make fuller use of target-language resources. Across all syntactic metrics, PEMT did not occupy an intermediate position between MT and HT. Instead, the post-edited text sample often introduced additional structural changes that exceeded those found in HT, particularly in word- and sequence-level crossing links as well as ASTrED scores. This suggests that PE in the literary domain may amplify syntactic divergence rather than merely converging toward HT patterns.

Such an outcome highlights the active role of post-editors. They intervene to correct and improve MT output, but their interventions do not necessarily replicate HT strategies. The result is a hybrid mode that reflects neither the syntactic conservatism of MT nor the stylistic creativity of HT. This observation resonates with research on post-editese (Castilho et al., 2019; Castilho and Resende, 2022), which emphasizes that PEMT carries its own linguistic fingerprints rather than serving as a mere compromise between MT and HT.

These findings are significant for both translation studies and MT research. For translation studies, they confirm that syntactic diversity is an important dimension of translation quality, especially in genres where stylistic variation is central. For MT developers, the results indicate that literary texts continue to expose weaknesses in current systems. Despite fluent outputs, structural conservatism remains a bottleneck, and PE seems to not simply solve the issue but rather create syntactic patterns of its own, even though the present sample size is too small to generalize. This points to the need for systems that can handle more flexible reordering and better accommodate the structural demands of creative texts.

Nevertheless, several limitations of the present study must be acknowledged.

This analysis was based on 24 texts, with only a single PEMT available, which is not representative of PE practice in the literary domain. Automatic word alignments were used throughout, enabling scalability but inevitably introducing noise, since manual alignments would provide greater precision. Furthermore, a small number of sentence pairs had to be excluded due to problematic parsing. In particular, ASTrED requires well-formed dependency trees with a single root, but some automatically parsed sentences produces malformed trees containing multiple roots, often as a result of parsing errors or irregular sentence structures. These pairs were filtered out prior to evaluation, ensuring the validity of the metric but slightly reducing the data available. Preprocessing steps also involved the deletion of misaligned or otherwise problematic sentences, which may have introduced further distortions in the data. Beyond these technical aspects, there are also conceptual limitations. An analysis based solely on syntactic structures cannot be equated with an overall assessment of translation quality. It rather captures one specific dimension of translational behavior. While syntax provides valuable insight into how structural patterns are changed or maintained across translation modes, other aspects, such as semantics, cohesion, and stylistic nuance, remain outside the scope of this investigation. In addition, the analysis is limited to a single language pair, Brazilian Portuguese and German, which restricts the generalizability of the findings to literary translation as a whole. Despite these constraints, the consistent patterns observed across multiple syntactic metrics show a clear trend that support the core findings.

Taken together, these findings highlight the complex position of PEMT in the spectrum between MT and HT. Rather than serving as a simple compromise, PEMT emerges as a distinct mode of translation with its own syntactic profile. At least in literary contexts, PE may diverge even further from both MT and HT. This underlines the need for further research on the role and value of PE in literary translation, where stylistic and structural creativity is central.

10. Conclusion and Future Work

This thesis investigated syntactic diversity in HT, MT, and PEMT of literary texts. Using a set of dependency-based metrics, including label changes, POS-tag changes, word and sequence crossing links, SACr, and ASTrED, the study examined to what extent PEMT aligns with MT or HT. The experiments revealed a consistent pattern that MT tends to remain structurally close to the ST, HT involves greater syntactic restructuring, and PEMT does not simply occupy a middle ground but instead emerges as a distinct mode of translation. In the case of literary texts, PEMT frequently introduced more syntactic divergence than either MT or HT, reflecting the hybrid and interventionist character of PE.

These findings contribute to translation studies by extending syntactic evaluation to the literary domain, which has received so far little attention compared to technical or institutional texts. They also suggest that PEMT cannot be assumed to replicate the creativity and restructuring strategies characteristic of HT, nor to preserve the source-oriented conservatism of MT. Instead, PEMT bears traces of both modes, while at the same time generating its own structural patterns. This has implications for the role of PE in literary contexts, where stylistic and syntactic choices are central to translation quality and reception.

Future work should expand on these results in several directions. A larger and more balanced corpus, including multiple PEMTs, would be necessary to validate the present observations and to better capture variation across genres, translators, and PE strategies. The precision of syntactic metrics could also be enhanced through the use of manual alignments or alternative alignment tools, reducing the dependency on noisy automatic outputs. Beyond this, it would be fruitful to investigate how syntactic diversity interacts with other dimensions of literary translation, such as narrative voice, character perspective, or stylistic register. Such work would not only refine our understanding of PEMT as a translation practice but would also situate it more firmly within the broader landscape of literary translation, where questions of creativity, interpretation, and reader experience remain central.

Bibliography

Lars Ahrenberg (2017). Comparing machine translation and human translation: A case study. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria. Association for Computational Linguistics, Shoumen, Bulgaria.

L. Aidoo, D.F. Silva, E. Fitz, R.R.M. Wasserman, N.H. Vieira, L.M. Schwarcz, E.K.F. Oliveira, P. da Luz-Moreira, V.A. Santos, R. Anderson, et al. (2013). *Lima Barreto: New Critical Perspectives*. New Critical Perspectives. Lexington Books.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR)*, San Diego, United States. Presented at ICLR 2015.

Mona Baker (1993). Corpus linguistics and translation studies implications and applications. In *Text and Technology*, pages 233–250. John Benjamins Publishing Company, The Netherlands.

Sabrina Baldo de Brébisson (2019). Comparison between automatic and human subtitling: A case study with game of thrones. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 1–10, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Satanjeev Banerjee and Alon Lavie (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Michele Banko and Robert C. Moore (2004). Part-of-speech tagging in context. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 556–561, Geneva, Switzerland. COLING.

Jean Boase-Beier (2011). *Critical Introduction To Translation Studies*. Continuum, London; New York.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi (2017). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Claudine Borg (2022). *Literary Translation In The Making : A Process-Oriented Perspective.* Routledge, New York, NY.

Oswald Campesato (2020). *Artificial intelligence, machine learning, and deep learning.* Mercury Learning and Information, Dulles, Virginia.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way (2017). Is neural machine translation the new state of the art? *Prague bulletin of mathematical linguistics*, 108(1):109–120.

Sheila Castilho, Natália Resende, and Ruslan Mitkov (2019). What influences the features of post-editese? a preliminary study. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 19–27, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Sheila Castilho and Natália Resende (2022). Post-editese in literary translations. *Information*, 13(2).

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors (2006). *Semi-supervised learning.* Adaptive computation and machine learning. MIT Press, Cambridge, Mass.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Joke Daems, Orphée De Clercq, and Lieve Macken (2017). Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia*, 16:89.

Giselle de Almeida and Sharon O'Brien (2010). Analysing post-editing performance: Correlations with years of translation experience. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Orphée de Clercq, Gert de Sutter, Rudy Loock, Bert Cappelle, and Koen Plevoets (2021). Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated french. *Translation Quarterly*, (101):21–45.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA.

Aleksei Dorkin and Kairit Sirts (2023). Comparison of current approaches to lemmatization: A case study in Estonian. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 280–285, Tórshavn, Faroe Islands. University of Tartu Library.

Chris Dyer, Victor Chahuneau, and Noah A. Smith (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Steffen Eger, Rüdiger Gleim, and Alexander Mehler (2016). Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1507–1513, Portorož, Slovenia. European Language Resources Association (ELRA).

Margot Fonteyne, Arda Tezcan, and Lieve Macken (2020). Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

William Frawley (1984). *Translation : literary, linguistic, and philosophical perspectives / edited by William Frawley.* University of Delaware Press, Newark.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 961–968, Sydney, Australia. Association for Computational Linguistics.

Martin Gellerstam (1986). Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster, editors (2009). *Learning machine translation*. Neural information processing series. MIT Press, Cambridge, Mass.

Richard Graham, editor (1999). *Machado de Assis*. University of Texas Press, New York, USA.

Ana Guerberof-Arenas and Antonio Toral (2020). The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins (2024). xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

James Luke Hadley, Kristiina Taivalkoski-Shilov, Carlos S. C. Teixeira, and Antonio Toral, editors (2022). *Using technologies for creative-text translation*. Routledge advances in translation and interpreting studies. Routledge, New York, NY.

Bettina Hiebl and Dagmar Gromann (2024). Comparative quality assessment of human and machine translation with best-worst scaling. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 507–536, Sheffield, UK. European Association for Machine Translation (EAMT).

Sepp Hochreiter and Jürgen Schmidhuber (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Florentina Hristea and Cornelia Caragea, editors (2022). *Natural Language Processing (NLP) and Machine Learning (ML)—Theory and Applications*. MDPI.

ISO 18587:2017 (2017). Translation services `Post-editing of machine translation output `Requirements. Iso, International Organization for Standardization, Geneva, CH.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Ridong Jiang, Rafael E. Banchs, and Haizhou Li (2016). Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27, Berlin, Germany. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Philipp Koehn (2010). *Statistical machine translation*. Cambridge University Press, Cambridge ; New York.

Philipp Koehn (2020). *Neural Machine Translation*. Cambridge University Press.

Waltraud Kolb (2023). Brazilian short prose in german: A study of literary post-editing. *Revista Tradumàtica*, 21(21):63–86.

Hans P Krings (2001). *Repairing texts : empirical investigations of machine translation ; post-editing processes*. Kent State Univ. Press.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Ting Liu, Chi-kiu Lo, Elizabeth Marshman, and Rebecca Knowles (2024). Evaluation briefs: Drawing on translation studies for human evaluation of MT. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 190–208, Chicago, USA. Association for Machine Translation in the Americas.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic (2024). The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the*

Association for Machine Translation in the Americas (Volume 2: Presentations), pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt (2014). Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):455–463.

Julie B Lovins (1968). *Development of a Stemming Algorithm.*

Lieve Macken, Bram Vanroy, Luca Desmet, and Arda Tezcan (2022). Literary translation as a three-stage process: machine translation, post-editing and revision. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 101–110, Ghent, Belgium. European Association for Machine Translation.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Evgeny Matusov (2019). The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Helena Moniz and Carla Parra Escartín, editors (2023). *Towards Responsible Machine Translation : Ethical and Legal Considerations in Machine Translation*, 1st ed. 2023 edition. Machine Translation: Technologies and Applications 4. Springer International Publishing Imprint: Springer, Cham.

Joss Moorkens (2025). *Automating translation.* Routledge introductions to translation and interpreting. Routldege, Abingdon, Oxon New York, NY.

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.

Kevin P Murphy (2012). *Machine learning : a probabilistic perspective.* Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts London, England.

Brian Nelson and Brigid Maher, editors (2013). *Perspectives on Literature and Translation: Creation, Circulation, Reception*. Routledge, New York.

Mara Nunziatini and Lena Marg (2020). Machine translation post-editing levels: Breaking away from the tradition and delivering a tailored service. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 309–318, Lisboa, Portugal. European Association for Machine Translation.

Franz Josef Och and Hermann Ney (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Robert Östling and Jörg Tiedemann (2016). Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung (2020). An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China. Association for Computational Linguistics.

Thierry Poibeau (2017). *Machine translation*. The MIT Press essential knowledge series. The MIT Press, Cambridge, Massachusetts.

Maja Popović (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović (2018). *Error Classification and Analysis for Machine Translation Quality Assessment*, pages 129–158. Springer International Publishing, Cham.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. (2018). Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21.

Gábor Recski and Fanni Kádár (2023). Language complexity in human and machine translation: a preliminary study. In *Proceedings of the International Conference HiT-IT 2023*, pages 268–281, Shoumen. Incoma Ltd.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Andrew Rothwell (2023). *Translation tools and technologies*, 1st edition edition. Routledge introductions to translation and interpreting. Routledge, Taylor Francis Group, London New York.

Andrew Rothwell, Andy Way, and Roy Youdale, editors (2023). *Computer-assisted literary translation*. Routledge advances in translation and interpreting studies. Routledge, New York, NY.

Aleksi Sahala and Krister Lindén (2023). A neural pipeline for POS-tagging and lemmatizing cuneiform languages. In *Proceedings of the Ancient Language Processing Workshop*, pages 203–212, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mehmet Şahin and Sabri Gürses (2019). Would MT kill creativity in literary retranslation? In *Proceedings of the Qualities of Literary Machine Translation*, pages 26–34, Dublin, Ireland. European Association for Machine Translation.

Elisabete Ranchhod Satoshi Sekine, editor (2009). *Named entities : recognition, classification, and use*. Benjamins current topics ; v. 19. John Benjamins Pub. Co., Amsterdam ; Philadelphia.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner (2024). Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Henry Schogt (1988). *Linguistics, Literary Analysis, and Literary Translation*. University of Toronto Press, Toronto.

Bernard Scott (2018). *Translation, Brains and the Computer: A Neurolinguistic Solution to Ambiguity and Complexity in Machine Translation*, 1st ed. 2018. edition, volume 2 of *Machine Translation: Technologies and Applications*. Springer International Publishing AG, Cham.

Anastassia Shaitarova, Anne Göhring, and Martin Volk (2023). Machine vs. human: Exploring syntax and lexicon in German translations, with a spotlight on anglicisms. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 215–227, Tórshavn, Faroe Islands. University of Tartu Library.

Isabel Slawik, Jan Niehues, and Alex Waibel (2015). Stripping adjectives: Integration techniques for selective stemming in SMT systems. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey. European Association for Machine Translation.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou (2021). Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maria Stasimioti and Vilelmini Sosoni (2020). Translation vs post-editing of NMT output: Insights from the English-Greek language pair. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 109–124, Virtual. Association for Machine Translation in the Americas.

Milan Straka, Jan Hajič, and Jana Straková (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Jana Straková, Milan Straka, and Jan Hajič (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Antonio Toral (2019). Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral and Andy Way (2015). Machine-assisted translation of literary text: A case study. *Translation Spaces*, 4(2):240–267.

Antonio Toral, Martijn Wieling, and Andy Way (2018). Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in digital humanities*, 5.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample (2023). Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning (2005). Morphological features help POS tagging of unknown words across language varieties. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Esther Gimeno Ugalde (2023). Carolina borges, alice leal, kathrin sartingen e melanie strasser (eds). contos do brasil. 200 anos de literatura brasileira/ erzählungen aus brasilien. 200 jahre brasilianische literatur, volumes i e ii. universität wien/brasilianische botschaft in wien, 2022. *Compendium (Lisboa)*, (4):127–130.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord (2020). UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Bram Vanroy, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken (2021). Metrics of syntactic equivalence to assess translation difficulty. In Michael Carl and Andy Way, editors, *Explorations in empirical translation process research*, volume 3 of *Machine Translation: Technologies and Applications*, pages 259–294. Springer International Publishing, Cham, Switzerland.

Bram Vanroy, Arda Tezcan, and Lieve Macken (2019). Predicting syntactic equivalence between source and target sentences. *Computational Linguistics in the Netherlands Journal*, 9:101–116.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lawrence Venuti (1995). *The Translator's Invisibility: A History of Translation*. Routledge, New York, NY.

Lawrence Venuti (2018). *The Translator's Invisibility: A History of Translation*, second edition. edition. Routledge, Taylor Francis Group, Abingdon, Oxon ; New York, NY.

Rob Voigt and Dan Jurafsky (2012). Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Lise Volkart and Pierrette Bouillon (2024). Post-editors as gatekeepers of lexical and syntactic diversity: Comparative analysis of human translation and post-editing in professional settings. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 387–395, Sheffield, UK. European Association for Machine Translation (EAMT).

Emma Wagner (1985). Post-editing systran – a challenge for commission translators. *Terminologie et Traduction*, 3:1–7.

Kelly Washbourne and Ben Van Wyke, editors (2019). *The Routledge handbook of literary translation*, 1st ed. edition. Routledge handbooks in translation and interpreting studies. Routledge, London.

Andy Way (2013). Traditional and emerging use-cases for machine translation. In *Proceedings of Translating and the Computer*, London. ASLIB.

Jonathan J. Webster and Chunyu Kit (1992). Tokenization as the initial phase in NLP. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.

Rebecca Webster, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems (2020). Gutenberg goes neural: Comparing features of dutch human translations with raw neural machine translation outputs in a corpus of english literary classics. *Informatics*, 7(3).

Chantal Wright (2016). *Literary translation*. Routledge translation guides. Routledge, New York, NY.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Roy Youdale (2020). *Using computers in the translation of literary style : challenges and opportunities*, 1st ed. edition. Routledge advances in translation and interpreting studies. Routledge, New York, NY.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Bryan Khelven da Silva Barbosa, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Ansu Berg, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Nina Böbel, Victoria Bobicev, Loïc Boizou, Johnatan Bonilla, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha

Burchardt, Carmen Cabeza, Natalia Cáceres Arandia, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Anila Çepani, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Claudine Chamoreau, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Netanel Dahan, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Ariani Di Felippo, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Ján Faryad, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, José María García-Miguel, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Gili Goldin, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Vladimir Gurevich, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Petter Hohle, Nick Howell, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre

Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Nelda Kote, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Emmanuelle Kowner, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Israel Landau, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Edita Luftiu, Arsenii Lukashevskyi, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Hilla Merhav, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Emilia Milano, Aaron Miller, Yael Minerbi, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Oggi Peeters, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci,

Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Ella Rabinovich, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Norton Trevisan Roman, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Paulette Roulon, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Rozana Rushiti, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Konstantinos Sampanis, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Gyu-Ho Shin, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Omer Strass, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Hakyung Sung, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Luigi Talamo, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Tarık Emre Tıraş, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Leonie Weissweiler, Maximilan Wendt,

Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Miriam Winkler, Shuly Wintner, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, Rayan Ziane, and Artūrs Znotiņš (2024). Universal dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Junru Zhou, Zuchao Li, and Hai Zhao (2020). Parsing all: Syntax and semantics, dependencies and spans. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4438–4449, Online. Association for Computational Linguistics.

Junru Zhou and Hai Zhao (2019). Head-Driven Phrase Structure Grammar parsing on Penn Treebank. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2396–2408, Florence, Italy. Association for Computational Linguistics.

A. Appendix

A.1. Supplementary Corpus Statistics

Table 16.: Evaluation metrics for MT and HT of Barreto Text 1: *O Pavilhão e a Pinel*

Metric	**MT**	**HT**
BLEU	31.89	–
TER	59.27	–
chrF	55.26	–
COMET	31.39	–
XCOMET	75.95	73.42

Table 17.: Corpus statistics for Barreto Text 2: *O nosso feminismo*

Metric	**ST**	**MT**	**HT**
Number of sentences	31	31	31
Total words	674	612	668
Unique lemmata	276	291	304
Characters	3745	4074	4353
TTR	0.494	0.574	0.549
Mean sentence length	21.74	19.74	21.55
Average length difference	–	2.65	2.23

Table 18.: Evaluation metrics for MT and HT of Barreto Text 2: *O nosso feminismo*

Metric	**MT**	**HT**
BLEU	29.13	–
TER	55.54	–
chrF	57.05	–
COMET	52.95	–
XCOMET	77.47	76.37

Table 19.: Corpus statistics for Barreto Text 3: *Não se zanguem*

Metric	**ST**	**MT**	**HT**
Number of sentences	12	12	12
Total words	318	300	306
Unique lemmata	160	175	175
Characters	1749	2072	2124
TTR	0.591	0.673	0.690
Mean sentence length	26.50	25.00	25.50
Average length difference	–	2.75	2.33

Table 20.: Evaluation metrics for MT and HT of Barreto Text 3: *Não se zanguem*

Metric	**MT**	**HT**
BLEU	26.14	–
TER	56.62	–
chrF	53.28	–
COMET	25.87	–
XCOMET	74.92	65.45

Table 21.: Evaluation metrics for MT and HT of Barreto Text 4: *A lei*

Metric	**MT**	**HT**
BLEU	30.79	–
TER	58.80	–
chrF	54.01	–
COMET	49.60	–
XCOMET	78.25	71.20

Table 58.: Evaluation metrics for MT and HT of Machado Text 1: *O Dicionário*

Metric	**MT**	**HT**
BLEU	18.43	–
TER	69.56	–
chrF	47.25	–
COMET	32.02	–
XCOMET	79.28	69.09

Table 22.: Corpus statistics for Barreto Text 5: *A volta*

Metric	**ST**	**MT**	**HT**
Number of sentences	10	10	10
Total words	218	226	217
Unique lemmata	126	133	135
Characters	1277	1530	1465
TTR	0.642	0.673	0.710
Mean sentence length	21.80	22.60	21.70
Average length difference	–	2.80	2.40

Table 23.: Evaluation metrics for MT and HT of Barreto Text 5: *A volta*

Metric	**MT**	**HT**
BLEU	35.38	–
TER	55.76	–
chrF	61.36	–
COMET	48.05	–
XCOMET	76.75	72.99

Table 24.: Corpus statistics for Barreto Text 6: *Nao as matem*

Metric	**ST**	**MT**	**HT**
Number of sentences	20	20	20
Total words	473	434	466
Unique lemmata	194	209	218
Characters	2541	2795	2939
TTR	0.518	0.599	0.584
Mean sentence length	23.65	21.70	23.30
Average length difference	–	2.45	1.65

Table 25.: Evaluation metrics for MT and HT of Barreto Text 6: *Nao as matem*

Metric	**MT**	**HT**
BLEU	38.63	–
TER	45.39	–
chrF	61.70	–
COMET	64.80	–
XCOMET	80.52	78.06

Table 26.: Corpus statistics for Barreto Text 7: *Exemplo a imitar*

Metric	**ST**	**MT**	**HT**
Number of sentences	18	18	18
Total words	306	290	295
Unique lemmata	147	153	164
Characters	1728	2001	2036
TTR	0.588	0.652	0.698
Mean sentence length	17.00	16.11	16.39
Average length difference	–	2.83	1.83

Table 27.: Evaluation metrics for MT and HT of Barreto Text 7: *Exemplo a imitar*

Metric	**MT**	**HT**
BLEU	27.47	–
TER	64.51	–
chrF	53.32	–
COMET	43.16	–
XCOMET	73.60	69.59

Table 28.: Corpus statistics for Barreto Text 8: *Uma lembrança*

Metric	**ST**	**MT**	**HT**
Number of sentences	21	21	21
Total words	312	282	282
Unique lemmata	162	158	162
Characters	1813	2010	1977
TTR	0.599	0.688	0.720
Mean sentence length	14.86	13.43	13.43
Average length difference	–	2.48	2.57

Table 29.: Evaluation metrics for MT and HT of Barreto Text 8: *Uma lembrança*

Metric	**MT**	**HT**
BLEU	44.15	–
TER	49.65	–
chrF	65.09	–
COMET	84.34	–
XCOMET	82.48	81.44

Table 30.: Corpus statistics for Barreto Text 9: *Cada raça tem um Calino*

Metric	**ST**	**MT**	**HT**
Number of sentences	83	83	83
Total words	1176	1146	1193
Unique lemmata	436	450	480
Characters	6454	7450	7897
TTR	0.459	0.499	0.503
Mean sentence length	14.17	13.81	14.37
Average length difference	–	2.02	2.24

Table 31.: Evaluation metrics for MT and HT of Barreto Text 9: *Cada raça tem um Calino*

Metric	**MT**	**HT**
BLEU	35.41	–
TER	50.38	–
chrF	61.02	–
COMET	57.08	–
XCOMET	79.01	77.60

Table 32.: Corpus statistics for Barreto Text 10: *Velhos apedidos e velhos anúncios*

Metric	**ST**	**MT**	**HT**
Number of sentences	100	100	100
Total words	1816	1773	1858
Unique lemmata	604	661	684
Characters	9833	11380	12007
TTR	0.425	0.468	0.459
Mean sentence length	18.16	17.73	18.58
Average length difference	–	2.85	2.55

Table 33.: Evaluation metrics for MT and HT of Barreto Text 10: *Velhos apedidos e velhos anúncios*

Metric	**MT**	**HT**
BLEU	29.31	–
TER	58.90	–
chrF	53.88	–
COMET	28.6	–
XCOMET	74.47	71.00

Table 34.: Corpus statistics for Barreto Text 11: *No primor da elegância*

Metric	**ST**	**MT**	**HT**
Number of sentences	132	132	132
Total words	1287	1279	1338
Unique lemmata	445	479	503
Characters	6917	7897	8173
TTR	0.454	0.472	0.468
Mean sentence length	9.75	9.69	10.14
Average length difference	–	1.72	1.83

Table 35.: Evaluation metrics for MT and HT of Barreto Text 11: *No primor da elegância*

Metric	**MT**	**HT**
BLEU	36.13	–
TER	49.74	–
chrF	56.57	–
COMET	49.51	–
XCOMET	73.34	72.80

Table 36.: Corpus statistics for Barreto Text 12: *Modas femininas e outras*

Metric	**ST**	**MT**	**HT**
Number of sentences	16	16	16
Total words	457	416	436
Unique lemmata	220	226	237
Characters	2490	2816	2926
TTR	0.562	0.642	0.635
Mean sentence length	28.56	26.00	27.25
Average length difference	–	3.56	3.12

Table 37.: Evaluation metrics for MT and HT of Barreto Text 12: *Modas femininas e outras*

Metric	**MT**	**HT**
BLEU	22.05	–
TER	63.43	–
chrF	51.08	–
COMET	33.66	–
XCOMET	72.36	66.83

Table 38.: Corpus statistics for Barreto Text 13: *Uma Partida de football*

Metric	**ST**	**MT**	**HT**
Number of sentences	16	16	16
Total words	308	273	283
Unique lemmata	156	163	167
Characters	1731	1921	1968
TTR	0.597	0.714	0.714
Mean sentence length	19.25	17.06	17.69
Average length difference	–	3.69	3.00

Table 39.: Evaluation metrics for MT and HT of Barreto Text 13: *Uma Partida de football*

Metric	**MT**	**HT**
BLEU	26.24	–
TER	59.01	–
chrF	55.25	–
COMET	38.66	–
XCOMET	80.14	76.12

Table 40.: Corpus statistics for Barreto Text 14: *As vaporosas*

Metric	**ST**	**MT**	**HT**
Number of sentences	19	19	19
Total words	250	250	262
Unique lemmata	119	127	138
Characters	1466	1616	1743
TTR	0.616	0.652	0.656
Mean sentence length	13.16	13.16	13.79
Average length difference	–	1.84	2.21

Table 41.: Evaluation metrics for MT and HT of Barreto Text 14: *As vaporosas*

Metric	**MT**	**HT**
BLEU	36.31	–
TER	48.48	–
chrF	57.73	–
COMET	46.32	–
XCOMET	75.86	68.69

Table 42.: Corpus statistics for Barreto Text 15: *Linhas de tiro*

Metric	**ST**	**MT**	**HT**
Number of sentences	24	24	24
Total words	464	404	430
Unique lemmata	208	199	212
Characters	2526	2564	2765
TTR	0.541	0.619	0.626
Mean sentence length	19.33	16.83	17.92
Average length difference	–	3.29	2.83

Table 43.: Evaluation metrics for MT and HT of Barreto Text 15: *Linhas de tiro*

Metric	**MT**	**HT**
BLEU	32.08	–
TER	53.04	–
chrF	59.04	–
COMET	47.96	–
XCOMET	70.36	67.75

Table 44.: Corpus statistics for Barreto Text 16: *Efeitos da lei valetudinária*

Metric	**ST**	**MT**	**HT**
Number of sentences	26	26	26
Total words	320	324	345
Unique lemmata	158	167	183
Characters	1771	2115	2233
TTR	0.591	0.648	0.655
Mean sentence length	12.31	12.46	13.27
Average length difference	–	1.92	2.58

Table 45.: Evaluation metrics for MT and HT of Barreto Text 16: *Efeitos da lei valetudinária*

Metric	**MT**	**HT**
BLEU	38.12	–
TER	46.22	–
chrF	63.47	–
COMET	68.44	–
XCOMET	79.47	76.28

Table 46.: Corpus statistics for Barreto Text 17: *Qualquer serve*

Metric	**ST**	**MT**	**HT**
Number of sentences	30	30	30
Total words	310	295	294
Unique lemmata	146	156	153
Characters	1649	1815	1840
TTR	0.587	0.664	0.643
Mean sentence length	10.33	9.83	9.80
Average length difference	–	1.17	1.63

Table 47.: Evaluation metrics for MT and HT of Barreto Text 17: *Qualquer serve*

Metric	**MT**	**HT**
BLEU	48.63	–
TER	40.68	–
chrF	65.73	–
COMET	76.13	–
XCOMET	80.16	77.08

Table 48.: Corpus statistics for Barreto Text 18: *Uma outra*

Metric	**ST**	**MT**	**HT**
Number of sentences	53	53	53
Total words	623	629	660
Unique lemmata	271	284	292
Characters	3351	3861	4039
TTR	0.554	0.571	0.564
Mean sentence length	11.75	11.87	12.45
Average length difference	–	1.64	1.94

Table 49.: Evaluation metrics for MT and HT of Barreto Text 18: *Uma outra*

Metric	**MT**	**HT**
BLEU	28.25	–
TER	57.08	–
chrF	52.61	–
COMET	43.18	–
XCOMET	73.37	68.75

Table 50.: Corpus statistics for Barreto Text 19: *Fala o corvo*

Metric	**ST**	**MT**	**HT**
Number of sentences	27	27	27
Total words	534	492	516
Unique lemmata	206	224	232
Characters	2918	3238	3458
TTR	0.491	0.577	0.572
Mean sentence length	19.78	18.22	19.11
Average length difference	–	2.56	2.15

Table 51.: Evaluation metrics for MT and HT of Barreto Text 19: *Fala o corvo*

Metric	**MT**	**HT**
BLEU	27.88	–
TER	55.95	–
chrF	54.91	–
COMET	48.28	–
XCOMET	75.08	67.65

Table 52.: Corpus statistics for Barreto Text 20: *Papel moeda*

Metric	**ST**	**MT**	**HT**
Number of sentences	19	19	19
Total words	428	396	413
Unique lemmata	187	201	217
Characters	2228	2521	2620
TTR	0.530	0.611	0.632
Mean sentence length	22.53	20.84	21.74
Average length difference	–	4.00	4.11

Table 53.: Evaluation metrics for MT and HT of Barreto Text 20: *Papel moeda*

Metric	**MT**	**HT**
BLEU	28.38	–
TER	60.19	–
chrF	53.77	–
COMET	52.51	–
XCOMET	78.53	74.54

Table 54.: Corpus statistics for Barreto Text 21: *Uma anedota*

Metric	**ST**	**MT**	**HT**
Number of sentences	30	30	30
Total words	199	208	228
Unique lemmata	109	113	120
Characters	994	1159	1262
TTR	0.653	0.673	0.667
Mean sentence length	6.63	6.93	7.60
Average length difference	–	1.43	1.83

Table 55.: Evaluation metrics for MT and HT of Barreto Text 21: *Uma anedota*

Metric	**MT**	**HT**
BLEU	46.61	–
TER	37.22	–
chrF	62.83	–
COMET	86.94	–
XCOMET	79.03	77.74

Table 56.: Corpus statistics for Barreto Text 22: *A questão dos telefones*

Metric	**ST**	**MT**	**HT**
Number of sentences	29	29	29
Total words	378	330	377
Unique lemmata	156	167	182
Characters	2003	2145	2353
TTR	0.495	0.606	0.610
Mean sentence length	13.03	11.38	13.00
Average length difference	–	2.48	1.38

Table 57.: Evaluation metrics for MT and HT of Barreto Text 22: *A questão dos telefones*

Metric	**MT**	**HT**
BLEU	29.58	–
TER	57.37	–
chrF	55.17	–
COMET	41.74	–
XCOMET	70.86	68.10

Table 59.: Corpus statistics for Machado Text 2: *A cartomante*

Metric	**ST**	**MT**	**HT**
Number of sentences	199	199	199
Total words	2717	2756	2729
Unique lemmata	726	785	883
Characters	14488	16935	17092
TTR	0.363	0.374	0.403
Mean sentence length	13.65	13.85	13.71
Average length difference	–	2.13	2.65

Table 60.: Evaluation metrics for MT and HT of Machado Text 2: *A cartomante*

Metric	**MT**	**HT**
BLEU	23.44	–
TER	68.36	–
chrF	48.00	–
COMET	34.86	–
XCOMET	75.39	67.47

Table 61.: Relative frequencies of dependency labels across modalities

Dep Label	ST	HT	MT	PEMT
punct	15.19	15.91	15.80	14.71
det	14.04	11.36	10.59	10.46
nsubj	5.45	9.23	9.34	9.15
case	10.80	7.00	7.01	8.36
advmod	5.42	7.27	6.98	7.66
obj	5.91	5.77	5.71	5.46
root	5.44	5.54	5.57	4.11
conj	4.74	4.60	4.83	5.18
obl	4.36	4.50	4.43	5.79
nmod	4.88	4.06	3.73	3.59
cc	3.56	3.27	3.59	3.36
amod	3.20	3.41	3.42	3.27
mark	2.96	2.60	2.89	2.15
cop	1.99	1.60	1.75	1.45
aux	0.46	1.82	2.00	1.96
acl	0.65	1.58	1.48	1.73
xcomp	1.61	0.92	0.96	1.35
advcl	1.45	0.99	1.13	0.89
det:poss	0.00	1.32	1.52	1.49
ccomp	1.00	1.00	1.04	0.89
obl:arg	0.00	1.16	1.03	1.63
appos	0.93	0.85	0.73	0.98
flat	0.02	0.85	0.96	1.26
parataxis	0.28	0.51	0.49	1.07
compound:prt	0.00	0.87	0.83	0.42
expl	0.76	0.36	0.37	0.19
acl:relcl	1.65	0.00	0.00	0.00
nummod	0.54	0.36	0.39	0.33
aux:pass	0.20	0.37	0.41	0.33
nsubj:pass	0.13	0.30	0.33	0.33
iobj	0.81	0.00	0.00	0.00
compound	0.05	0.24	0.27	0.23
fixed	0.71	0.02	0.02	0.00
flat:name	0.44	0.00	0.00	0.00
obl:agent	0.11	0.06	0.10	0.14
csubj	0.15	0.11	0.12	0.00
dep	0.00	0.13	0.11	0.05
nmod:poss	0.00	0.05	0.03	0.05
discourse	0.08	0.00	0.00	0.00
csubj:pass	0.00	0.02	0.03	0.00
expl:pv	0.00	0.02	0.02	0.00
flat:foreign	0.02	0.00	0.00	0.00
vocative	0.01	0.00	0.00	0.00

Table 62.: Relative frequencies of POS-tags across modalities

POS Tag	**ST**	**HT**	**MT**	**PEMT**
NOUN	16.75	15.90	15.24	14.15
PUNCT	15.20	15.95	15.83	14.71
DET	14.22	13.69	13.09	12.75
VERB	11.89	10.08	10.25	10.22
PRON	7.39	10.19	10.54	12.65
ADP	11.06	7.86	7.82	8.64
ADV	5.76	5.93	5.57	6.07
ADJ	4.58	5.94	6.15	6.35
AUX	2.77	4.01	4.43	4.01
CCONJ	3.55	3.25	3.55	3.36
PROPN	2.65	3.10	3.17	3.59
SCONJ	3.01	1.59	1.80	1.45
PART	0.00	1.67	1.79	1.59
NUM	1.01	0.63	0.62	0.42
X	0.03	0.15	0.13	0.00
INTJ	0.13	0.05	0.04	0.05