Entre rimas e algoritmos: uma investigação sobre tradução automática poética

Beatriz Ribeiro Borges

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Beatriz Ribeiro Borges

Entre rimas e algoritmos: uma investigação sobre tradução automática poética

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Elaine Ribeiro de Faria Paiva
Coorientador: Paulo Henrique Ribeiro Gabriel

Uberlândia
2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

B732
2026

Borges, Beatriz Ribeiro, 1993-
Entre rimas e algoritmos: uma investigação sobre tradução automática poética [recurso eletrônico] / Beatriz Ribeiro Borges. - 2026.

Orientadora: Elaine Ribeiro de Faria Paiva.
Coorientadora: Paulo Henrique Ribeiro Gabriel.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2026.110
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Paiva, Elaine Ribeiro de Faria,1980-, (Orient.). II. Gabriel, Paulo Henrique Ribeiro,1984-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG,
CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 06/2026, PPGCO				
Data:	11 de Fevereiro de 2026	Hora de início:	09:10	Hora de encerramento:	10:50
Matrícula do Discente:	12412CCP001				
Nome do Discente:	Beatriz Ribeiro Borges				
Título do Trabalho:	Entre rimas e algoritmos: uma investigação sobre tradução automática poética				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	-----------------				

Reuniu-se por presencialmente, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Paulo Henrique Ribeiro Gabriel (Coorientador) - FACOM/UFU, Murillo Guimarães Carneiro - FACOM/UFU, Ricardo Marcondes Marcacini - ICMC-USP e Elaine Ribeiro de Faria Paiva - FACOM/UFU, orientadora do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Todos os membros da banca e o aluno(a) participaram da cidade de Uberlândia.

Iniciando os trabalhos o(a) presidente da mesa, Prof. Dr. Elaine Ribeiro de Faria Paiva, apresentou a Comissão Examinadora e o(a) candidato(a), agradeceu a presença do público, e concedeu ao(à) Discente a palavra para a exposição do seu trabalho.

A seguir o senhora presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato(a):

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Elaine Ribeiro de Faria Paiva**, **Professor(a) do Magistério Superior**, em 11/02/2026, às 14:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Paulo Henrique Ribeiro Gabriel**, **Professor(a) do Magistério Superior**, em 11/02/2026, às 14:54, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **RICARDO MARCONDES MARCACINI**, **Usuário Externo**, em 11/02/2026, às 15:16, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Murillo Guimarães Carneiro**, **Professor(a) do Magistério Superior**, em 11/02/2026, às 15:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **7017872** e o código CRC **220B54E2**.

Referência: Processo nº 23117.004478/2026-44 SEI nº 7017872

Agradecimentos

Agradeço, primeiramente, à minha mãe, Sueli, pelo amor e apoio incondicional, pela alegria sincera a cada uma de minhas conquistas, pelo orgulho visível em cada passo meu e por me ensinar, desde sempre, a ter resiliência.

Ao Gustavo, pelo incentivo constante, pelo apoio, pela paciência em repassar comigo cada cenário, por comemorar comigo cada conquista e planejar os próximos passos sempre me incentivando a acreditar que consigo.

À minha orientadora, Profa. Dra. Elaine Ribeiro de Faria Paiva, e ao meu coorientador, Prof. Dr. Paulo Henrique Ribeiro Gabriel, pelas orientações cuidadosas e pelas valiosas contribuições intelectuais, mas também pelo apoio nesse caminho acadêmico e, acima de tudo, por aceitarem se aventurar comigo entre os campos da poesia.

Aos professores da Faculdade de Computação e ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Uberlândia, pelo apoio, pelas disciplinas ministradas, conhecimento compartilhado e pela infraestrutura e suporte oferecidos.

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

“Numa meia-noite agreste, quando eu lia, lento e triste,
Vagos curiosos tomos de ciências ancestrais,
E já quase adormecia, ouvi o que parecia
O som de alguém que batia levemente a meus umbrais.
‘Uma visita’, eu me disse, ‘está batendo a meus umbrais.
É só isto, e nada mais.’ ”
(Poe, 2017)

Resumo

A tradução poética é uma tarefa complexa que vai além da simples transferência semântica, exigindo a preservação de ritmo, rima, imagens, elementos estilísticos e nuances culturais. Apesar dos avanços na tradução automática, esses sistemas ainda apresentam limitações em contextos literários e poéticos. Este trabalho investiga as capacidades de modelos de tradução automática (TA) em seis pares de línguas (Português–Inglês, Inglês–Português, Português–Francês, Francês–Português, Inglês–Francês e Francês–Inglês), comparando modelos de tradução especializados, como Google Translate, MarianMT, mBART e OpenNMT usando Rede Recorrente Neural (RNN), incluindo fine-tuning com corpora de poemas e letras de música, com grandes modelos de linguagem (LLMs) como ChatGPT e Maritaca AI. Foi implementado um pipeline de avaliação em três fases, combinando (i) métricas automáticas (BLEU, METEOR, BERTScore, BARTScore) para medir similaridade lexical e semântica, (ii) modelagem de tópicos com BERTopic para avaliar a preservação temática, e (iii) avaliação humana especializada focada em estrutura poética, estilo, fluência e significado. Os resultados indicam que os LLMs e o Google Translate, superam consistentemente os modelos tradicionais de TA, enquanto o OpenNMT (RNN) apresentou desempenho inferior. O uso do prompt influenciou o desempenho dos LLMs: ChatGPT obteve maior fidelidade semântica com prompts gerais, enquanto Maritaca AI integrou melhor instruções poéticas explícitas em português. A avaliação humana confirmou melhores pontuações para LLMs em fluência e preservação do significado, embora todos os modelos tenham dificuldades em estrutura e estilo Poético. A modelagem de tópicos confirmou que esses modelos preservam melhor a consistência temática, muitas vezes alinhando-se mais às traduções humanas de referência do que aos poemas originais. Os efeitos do fine-tuning foram dependentes da arquitetura: mBART se beneficiou de poemas e letras de música, ao contrário do MarianMT e OpenNMT (RNN).

Palavras-chave: Tradução de Poesia. Tradução Automática Neural. Modelos de Linguagem de Grande Escala. Processamento de Linguagem Natural.

Between Rhymes and Algorithms: An Investigation into Poetic Machine Translation

Beatriz Ribeiro Borges

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2025

Abstract

Poetry translation is a complex task that extends beyond semantic transfer, requiring the preservation of rhythm, rhyme, imagery, stylistic elements, and cultural nuances. Despite advances in neural machine translation, these systems still show limitations in literary and poetic contexts. This work investigates the capabilities of automatic translation systems across six language pairs (Portuguese–English, English–Portuguese, Portuguese–French, French–Portuguese, English–French, and French–English), comparing specialized machine translation (MT) models, such as Google Translate, MarianMT, mBART, and OpenNMT (RNN), including fine-tuned versions trained on poetic and music lyrics corpora, with large language models (LLMs) such as ChatGPT and Maritaca AI. A three-phase evaluation pipeline was implemented, combining (i) automatic metrics (BLEU, METEOR, BERTScore, BARTScore) to assess lexical and semantic similarity, (ii) topic modeling with BERTopic to evaluate thematic preservation, and (iii) expert human evaluation focusing on poetic structure, style, fluency, and meaning. Results indicate that LLMs, and also Google Translate, consistently outperform traditional MT models, while OpenNMT with Recurrent Neural Network (RNN) performed worst. Prompt design influenced LLM performance: ChatGPT favored general prompts for higher semantic fidelity, whereas Maritaca AI better integrated explicit poetic instructions in portuguese. Human evaluation confirmed superior scores for LLMs in fluency and meaning preservation, though all models struggled with poetic Structure and stylistic. Topic modeling showed that top-performing models better preserved thematic consistency, often aligning more closely with human reference translations than with original poems. Fine-tuning effects were architecture-dependent: mBART benefited from poems and song lyrics, while MarianMT and OpenNMT (RNN) showed limited gains.

Keywords: Poetry Translation. Neural Machine Translation. Large Language Models. Natural Language Processing..

List of Figures

List of Tables

Acronyms list

BLEU Bilingual Evaluation Understudy

c-TF-IDF Class-based Term Frequency-Inverse Document Frequency

EN-FR English-French

EN-PT English-Portuguese

FR-EN French-English

FR-PT French-Portuguese

LLMs Large Language Models

MT Machine Translation

METEOR Metric for Evaluation of Translation with Explicit Ordering

PT-EN Portuguese-English

PT-FR Portuguese-French

Contents

Chapter 1

Introduction

1.1 Contextualization and Motivation

Translation is a complex process that functions as a bridge between languages and cultures, playing a crucial role in cultural transmission (OZTURK, 2024). Present in various aspects of daily life, from technical manuals to literary works, it goes beyond mere word substitution, requiring the interpretation of cultural nuances, implicit meanings, and contextual specificities (BRITTO, 2012).

Among the various types of translation, literary translation is essential to preserve cultural diversity, offering access to different ways of thinking and expressing across the world (CANDIDO, 2006). It also stands out due to the unique challenges it poses. More than just converting meanings between languages, it requires careful attention to the essence of the original work, considering its historical, social, and cultural context to preserve its authenticity. Within this field, poetic translation stands out for its complexity, as it demands balancing fidelity to the original with creativity to recreate crucial elements such as rhythm, rhyme, metaphors, and style (JONES, 2011).

Given the vast number of poetic texts in multiple languages, human translation of all works is impractical. Automatic translation provides a viable solution to broaden access to literature and disseminate cultural knowledge (CHATZIKOUMI, 2020). While Machine Translation (MT) systems can process large volumes efficiently, they still struggle with the semantic, syntactic, and stylistic challenges of poetry (CONSTANTINE, 2019).

Neural network–based MT has achieved significant advancements, yet important challenges remain, particularly in the translation of poetry. These include preserving poetic structure (rhythm, meter, rhyme), conveying figurative language, and cultural nuances. Within neural MT, the Transformer architecture is considered state of the art in recent evaluations (WANG et al., 2019), leading recent research to focus on specialized translation models and generative models.

As Large Language Models (LLMs) continue to evolve, MT has reached a new level. Unlike traditional models, LLMs aim to approach the human translation process, taking into account contextual and stylistic particularities. This not only improves technical translation, but also opens new avenues for poetic translation, a field that, despite being challenging, holds great potential for innovation and development (HE et al., 2024).

Despite significant advances in studies on neural MT, important gaps remain to be explored. In the field of neural networks, the Transformer architecture is considered state-of-the-art in recent evaluations of MT (WANG et al., 2019). Consequently, recent works have focused on specialized translation models, such as mBART or MarianMT, as well as on generative models like ChatGPT. However, few studies have conducted direct comparisons between these two approaches – generative models and specialized translation systems – limiting a deeper understanding of their potential, particularly in literary contexts and in Latin-based languages such as Portuguese and French.

Moreover, poetic translation constitutes a particularly demanding test bed for machine translation, as it extends beyond semantic adequacy and requires sensitivity to formal structure and stylistic choices. As a result, it provides a rigorous evaluation setting for both traditional MT models and LLMs.

An analysis combining these different types of models, using multiple evaluation metrics, is necessary to examine the convergence in results and to identify the strengths and weaknesses of each approach. Another motivation of this work lies in the insufficiency of existing evaluation methodologies for poetic translation. Traditional automatic metrics, such as Bilingual Evaluation Understudy (BLEU) (PAPINENI et al., 2002), Metric for Evaluation of Translation with Explicit Ordering (METEOR) (BANERJEE; LAVIE, 2005), and BERTScore (ZHANG et al., 2020), effectively measure lexical and semantic overlap, but are fundamentally insensitive to elements that define poetry. They fail to capture the preservation of theme, rhythm, rhyme, style, or aesthetic quality, resulting in incomplete and potentially misleading assessments of translation quality. This highlights the need for a more holistic evaluation approach that integrates quantitative precision with the qualitative sensitivity of human analysis.

The introduction of topic modeling as a tool to verify thematic consistency between the original text and its translation addresses another gap: ensuring that the poem's theme is preserved regardless of how it is translated. Based on these considerations, this study aims to evaluate the effectiveness of MT tools in poetry translation, seeking to understand the extent to which these systems can produce high-quality translations of the original poems.

To achieve this, a three-phase pipeline was established, combining automatic metrics, topic modeling, and human evaluation to compare the advantages and limitations of different translation systems. To further investigate the potential for improvement, we perform fine-tuning of selected models using a corpus composed of poems and an extended dataset that also includes song lyrics.

1.2 Objectives

The main objective of this work is to propose and apply a multi-perspective evaluation strategy capable of analyzing different dimensions of the MT of poetry. The specific objectives are:

- ❏ Analyze the main dimensions involved in the translation of poetic texts, including semantic quality, thematic coherence, and poetic structure.
- ❏ Compare the performance of MT models, including mBART, MarianMT, Google Translate, ChatGPT, Maritaca AI, and OpenNMT (RNN), across different language pairs (Portuguese, French, and English), using the proposed three-phase pipeline to obtain a comprehensive evaluation.
- ❏ Investigate the impact of fine-tuning with corpora of poems and song lyrics on poetic translation quality.

For this purpose, a three-phase evaluation framework was proposed, combining automatic metrics to assess lexical and semantic quality, topic modeling to examine thematic preservation, and expert human judgment to evaluate poetic structure, style, and meaning.

1.3 Hypothesis

This study has the following hypotheses:

- ❏ H_1: Large Language Models achieve higher performance in poetic translation than traditional MT systems, considering automatic metrics, topic modeling–based thematic analysis, and human evaluations of poetic quality.
- ❏ H_2: Fine-tuning translation models with poetry and song lyrics leads to improvements in translation scores.

The hypothesis will be evaluated using a three-phase pipeline that combines automatic metrics, topic modeling, and a structured human evaluation. Additionally, fine-tuning strategies with a corpus of poems and song lyrics will be explored to investigate potential improvements in capturing stylistic nuances.

1.4 Contributions

This study makes the following contributions:

- ❏ A comprehensive evaluation framework combining automatic metrics, topic-based analysis, and human expert assessment for poetry translation.
- ❏ Empirical insights into the performance of various MT systems, including specialized neural models, LLMs, and traditional RNN-based architectures.
- ❏ A methodological contribution by applying topic modeling to assess thematic preservation in translated poetry.
- ❏ An evaluation of fine-tuning strategies leveraging poems and song lyrics to enhance domain-specific translation performance.
- ❏ A multilingual perspective, evaluating Latin-based languages that are underrepresented in existing research.
- ❏ A structured human evaluation questionnaire for literary translation that can serve as a reference in future studies.
- ❏ The creation of a multilingual poetry dataset comprising six translation language pairs (FR-EN, FR-PT, EN-FR, EN-PT, PT-FR, PT-EN) with the original poems and its human reference translations (BORGES, 2025).
- ❏ Public release of the source code to support transparency and reproducibility[1].

1.5 Dissertation Organization

This dissertation is organized as follows. Chapter 2 presents the main concepts that will be used throughout the work, including MT, poetic translation, evaluation metrics, and topic modeling. Chapter 3 reviews the state-of-the-art works related to automatic translation of poetry and neural MT models. Chapter 4 details the development of the proposed methodology, including the datasets, preprocessing steps, translation models, fine-tuning strategies, and the three-phase evaluation pipeline. Chapter 5 presents the experimental setup, reports the results of automatic metrics, topic modeling, and human evaluation, and provides an analysis of these results. Chapter 6 summarizes the main conclusions of this study and discusses possible directions for future research. Finally, the appendices are presented after the conclusion.

[1] <https://github.com/biarborges/traducaoPoemasLLM>

Chapter 2

Fundamentals

The purpose of this chapter is to establish the theoretical foundation for understanding and developing the study, addressing the main concepts and techniques involved. Section 2.1, Translation, provides a brief overview of translation as a process of linguistic and cultural mediation, followed by a discussion of Literary Translation and, more specifically, Poetic Translation, which is the focus of this study, highlighting its unique characteristics. Section 2.2, Machine Translation, outlines the evolution of translation technologies, from rule-based and statistical approaches to modern neural architectures. Section 2.3, Specialized Machine Translation Models and LLMs, introduces the specialized machine translation systems and large language models considered in this study, including mBART, MarianMT, OpenNMT with RNN, Google Translate, ChatGPT, and Maritaca AI. Section 2.4, Evaluation Metrics, describes the main quantitative measures employed (BLEU, METEOR, and BERTScore) and presents the qualitative criteria applied by human experts. Section 2.5, Topic Modeling, introduces thematic analysis, which is used in this study as a complementary evaluation method, with particular emphasis on the BERTopic approach. Finally, Section 2.6 presents the final considerations of the chapter.

2.1 Translation

Translation is an essential activity in any society that interacts with other cultures and languages, as much of the information and materials consumed, from technical manuals to religious texts, undergo some form of translation (BRITTO, 2012). Far from being a mere mechanical substitution of words from one language into another, it constitutes a complex process of intercultural and linguistic mediation, encompassing interpretation, adaptation, and negotiation. Translators operate within intricate cultural and linguistic contexts, where they must balance not only linguistic equivalence but also cultural norms, idiomatic expressions, worldviews, and situational factors. Effective intercultural mediation often requires rearticulating meanings for audiences who do not share the source culture, which can involve reworking cultural references, adjusting pragmatic implications, and selecting strategies of domestication or foreignization (LIDDICOAT, 2016).

In addition to these interpretive challenges, translators often rely on guiding principles to navigate their work. For example, decisions about literal versus free translation involve choosing between preserving the exact words and structures of the source text (literal) or prioritizing meaning, tone, and readability in the target language (free). Choices regarding domestication versus foreignization determine whether the text is adapted to the target culture or retains elements of the source culture (VENUTI, 1995). Thus, translation emerges not merely as a linguistic operation, but as a situated and interpretive act, shaped by cultural, ideological, and contextual factors that influence the production of meaning.

2.1.1 Poetic Translation

Translating literary texts is an inherently challenging task that requires not only linguistic competence but also advanced cognitive and creative skills. Translators must interpret ambiguities, explore cultural connotations, and capture underlying symbolism, exercising sensitivity to literary language in order to recreate the work of art in the target language. This involves balancing fidelity to the original text with the need for creative adaptation to preserve its essence (BOASE-BEIER; FISHER; FURUKAWA, 2018).

The complexity increases when it comes to translating poetry, due to the unique characteristics of these literary texts, such as rhyme, wordplay, figures of speech, and rhythm, which often have no direct equivalents in other languages. These elements make the process of transferring meaning between languages challenging even for experienced professionals. Any error in interpreting the source text can distort the aesthetic and cultural aspects of the work, compromising its integrity and artistic impact (MADKOUR, 2016).

Translators in this domain act as co-creators, shaping tone, preserving artistic effect, and making difficult choices about which aspects to prioritize when trade-offs are unavoidable. Poetic translation, therefore, requires a careful balance between formal fidelity and semantic accuracy. Translators must decide whether to prioritize form or content, how to render stylistic features, and whether to domesticate or foreignize cultural references, all while preserving the poetic effect in the target language (PARKS, 2007). Consequently, a solid understanding of the fundamental principles of poetic literature is essential for producing high-quality translations. As highlighted by (GOLDSTEIN, 2005), the key foundations of poetic literature include:

❑ **Rhythm:** Rhythm in poetry is the pattern of sounds and beats created by meter, repetition, and stylistic variations, giving each poem its musicality. For example, in Vinicius de Moraes' children's poem "O sapo não lava o pé":

> "O sapo não lava o pé,
> Não lava porque não quer.
> Ele mora lá na lagoa,
> Não lava o pé porque não quer."

The rhythm in this excerpt is marked by the repetition of sounds and the regular cadence of the lines. Each line has approximately the same number of syllables, creating a sense of musicality and movement.

❑ **Meter:** Meter in poetry can be analyzed in two main ways through scansion (the breakdown of lines): the quantitative system and the syllabic-accentual system. In the quantitative system, used in classical Greek and Latin poetry, meter is defined by the alternation of long and short syllables. In the syllabic-accentual system, predominant in languages like Portuguese, the count of syllables is combined with the placement of stressed syllables, creating a rhythmic pattern.

Examples:

– **Quantitative-like (regular metrical pattern):** In English poetry, for instance, the opening lines of Shakespeare's *Sonnet 18*:

> "Shall I compare thee to a summer's day?
> Thou art more lovely and more temperate."

In this excerpt, each line follows an iambic pentameter: ten syllables with alternating unstressed and stressed syllables. The pattern can be represented as follows (u = unstressed, s = stressed):

> Line 1: Line 1: u – s – u – s – u – s –
> Shall I com- | pare thee to | a sum- | mer's day?

> Line 2: Line 1: u – s – u – s – u – s –
> Thou art more | love-ly and | more tem- | pe-rate

This regular pattern creates a rhythmic cadence that guides the reading and enhances the musicality of the poem. Minor natural variations may occur in pronunciation, but the overall structure preserves the iambic pentameter, producing a flowing and melodious effect.

– **Syllabic-accentual:** In Portuguese poetry, meter is often based on the syllabic-accentual system, which combines the number of syllables in a line with the placement of stressed syllables. One of the most common forms is the decasyllable (verso decassílabo), frequently used in sonnets. For example, in Luís de Camões' Sonnet 11:

> "Alma minha gentil, que te partiste
> Tão cedo desta vida, descontente,"

Each line consists of ten poetic syllables, with stresses typically placed on the 6th and 10th syllables, and the syllable count extends only up to the last stressed syllable. The scansion can be represented as:

> Al-ma mi-nha gen-til, que te par-tis
> Tão ce-do des-ta vi-da, des-con-ten

This regular distribution of stresses and syllables produces a measured cadence, typical of classical Portuguese verse, and contributes to the solemn and melodic tone characteristic of Camões' sonnets.

❑ **Verse:** Verses are the lines that structure a poem and can be classified according to their adherence to metric and rhythmic rules. Regular verses follow strict metric norms and maintain consistent rhymes, while blank verses preserve the meter but do not rhyme. Free verses, typical of Modernism[1], dispense fixed rules of meter, stress position, or rhyme, favoring expressive freedom and creative flow.

❑ **Stanza:** A stanza is a grouping of verses that forms a unit within a poem, visually separated by blank spaces and often marked by sonic elements such as rhyme. Stanza forms vary widely: they can be short, such as couplets (two verses), tercets (three verses), quatrains (four verses), or long, like octaves (eight verses) or more.

❑ **Rhyme:** Rhyme is the repetition of similar sounds that establishes a sonic relationship between words. This repetition can occur at the end of lines, within a single line, or in varying positions, depending on the poet's stylistic intent. Rhymes play a crucial role in the musicality of a poem, helping to create sound patterns, emphasize keywords, and reinforce the rhythm of the composition. The following are a few examples of common rhyme types.

Types of rhymes:

[1] Modernism was an artistic movement of the 20th century that brought together a variety of artistic groups with an innovative approach in the arts, breaking traditions and freeing from the past. Source: (VELLOSO, 2010)

- **According to the position:** Rhyme may occur in different positions within a verse. When the rhyme appears inside the same line, it is referred to as internal rhyme; when it occurs at the end of lines, it is known as end rhyme. Example: Haiku (a traditional Japanese three-line poem) "O Pensamento" by Guilherme de Almeida:

  "O ar. A folha. A fuga.
  No lago, um círculo vago.
  No rosto, uma ruga."

  In this haiku, "lago" and "vago" form an internal rhyme within the second line, while "fuga" and "ruga" create an end rhyme connecting the first and third lines.

- **Position in the stanza:** The arrangement of lines within a stanza can follow different organizational patterns, often associated with rhyme schemes. Common configurations include structures such as crossed rhyme (ABAB), paired (AABB), interpolated (ABBA) or mixed (combines different rhyme schemes within the same poem or stanza) which determine how lines relate to one another both sonically and structurally. Below is an example of a paired rhyme scheme:

  "Aos que me chamam de deputado (A)
  Quando nem mesmo sou jurado, (A)
  Aos que, de bons, se babam: mestre! (B)
  Inda se escrevo o que não preste." (B)

- **Rhyme richness:** Rhyme richness refers to the phonetic variety between rhyming words. It is classified as follows:

  - **Poor:** Rhymes in which the words share the same grammatical class. Example (excerpt from the poem "Quando Ela Fala" by Machado de Assis):

    Quando ela fala, parece
    Que a voz da brisa se cala;
    Talvez um anjo emudece
    Quando ela fala.

    In this example "fala" and "cala", and "parece" and "emudece" are verbs.

- * **Rich:** Rhymes in which the words belong to different grammatical classes. Example (excerpt from the poem “Relíquia Íntima” by Machado de Assis):

  Ilustríssimo, caro e velho amigo,
  Saberás que, por um motivo urgente,
  Na quinta-feira, nove do corrente,
  Preciso muito de falar contigo.

  In this example “amigo” is a noun and “contigo” is a pronoun; and “urgente” is an adjective and “corrente” is a noun.

- ❏ **Semantic Level:** The semantic level refers to the overall meaning of a poem, which emerges from the interaction of linguistic choices and figures of speech that create specific poetic effects. Figures of words or semantics involve modifications or extensions of the meaning of words, such as metaphors, and metonymy. Figures of thought operate at the conceptual or cognitive level, shaping ideas and interpretations through devices like hyperbole, and irony. Figures of syntax or construction affect the grammatical and structural arrangement of words and phrases, including ellipsis (the omission of a term that can be inferred from the context), and hyperbaton (it alters the normal word order), influencing the rhythm, emphasis, and flow of the poem. Finally, figures of sound focus on phonetic elements, such as alliteration (the repetition of consonant sounds), onomatopoeia, and repetition, enhancing musicality, and cohesion.

- ❏ **Lexical Level:** The lexical level of a poem involves analyzing the vocabulary used, identifying whether the language is formal, literary, or colloquial, which can reflect the author’s style and aesthetic intent.

- ❏ **Syntactic Level:** The syntactic level analyzes how sentences and clauses are organized within a poem, observing, for example, the type of punctuation (short or long sentences), and the presence of parallelisms (similar structures in different lines), which can create particular effects of meaning.

- ❏ **Enjambment:** Enjambment is a poetic device in which the meaning and syntactic structure of a line are completed only in the following line. Although the line has a complete metric structure, its meaning and syntax remain incomplete, generating a tension between sound, syntax, and sense. Example from Carlos Drummond de Andrade, “Quadrilha”:

  João amava Teresa que amava Raimundo
  que amava Maria que amava Joaquim
  que amava Lili que não amava ninguém...

> In this excerpt, the syntactic and semantic sense of each clause continues into the next line, creating a flowing, interconnected structure and highlighting the circularity and tension in the poem's theme of unrequited love.

Beyond its formal characteristics, poetry poses specific challenges for translation due to the frequent interplay between sound, meaning, and structure. Elements such as rhyme, meter, and wordplay are often language-specific and cannot be transferred directly without loss or transformation. For instance, rhyming patterns that are easily achievable in one language may require semantic compromises in another, forcing the translator to choose between preserving meaning or maintaining formal structure (JONES, 2011). The following stanza from Carlos Drummond de Andrade's poem "José" exemplifies the challenges of translation, accompanied by Len Sousa's translation.

> E agora, José?
> A festa acabou,
> a luz apagou,
> o povo sumiu,
> a noite esfriou,
> e agora, José?
> e agora, você?
> você que é sem nome,
> que zomba dos outros,
> você que faz versos,
> que ama, protesta?
> e agora, José?

> What now, José?
> The party's over,
> the lights are off,
> the crowd's gone,
> the night's gone cold,
> what now, José?
> what now, you?
> you without a name,
> who mocks the others,
> you who write poetry
> who love, protest?
> what now, José?

Based on the concepts of poetic structure, it's possible to examine the difficulty of translating a literary poetic work. The translation of these verses loses the rhyme and, consequently, the repetition present in the original. In Portuguese, there is a vowel echo that is lost in English.

In the line "Você que é sem nome, que zomba dos outros," the repeated /s/ sounds are prominent in the original, but this alliteration is absent in the English version, "you without a name, who mocks the others."

In Portuguese, "José" functions as a generic name, representing the "everyman." The question "E agora, José?" has become an idiomatic expression in Brazil to describe a dead-end or an existential impasse.

The issue is that, for an English-speaking reader, "José" may evoke a specific ethnicity that is not intended in the original. The translator must decide whether to retain the name (preserving the author's original identity) or adapt it to something like "What now, Jack?" to convey the sense of an "everyman," a choice that would alter the work.

2.2 Machine Translation

Automatic Translation or Machine Translation, is the process that enables the translation of texts from one natural language (source language) into another (target language) in an automated manner (RUSSELL; NORVIG, 2013). This field is part of a broader research context in Natural Language Processing, which is grounded in Computational Linguistics and Artificial Intelligence. Research in MT is closely related to these disciplines, as it adopts and applies both theoretical perspectives and operational techniques to translation processes (HUTCHINS; SOMERS, 1992).

This field has become increasingly relevant, reflecting the growing demand for automatic translations. However, the idea of translating texts automatically dates back to the 1950s. After World War II, at Georgetown University, the French-American linguist Leon Dostert collaborated with the *International Business Machines Corporation* (IBM) on a project that led to the first public demonstration of a MT system in January 1954. In this presentation, Russian sentences were translated into English using a limited vocabulary of 250 words and only six grammatical rules, introducing what became known as Rule-Based MT (HUTCHINS; SOMERS, 1992).

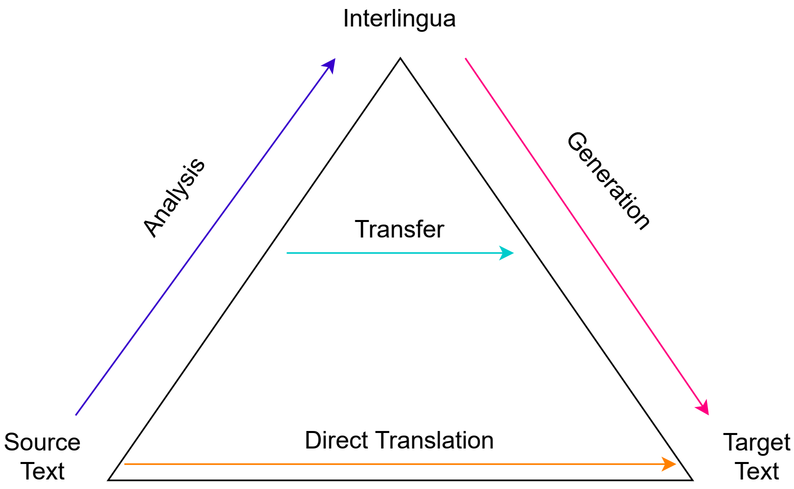
Since then, MT has evolved significantly, with different approaches available for automatic translation. Among the simplest methods is direct translation, which operates word by word or by short sequences of words. In contrast, the most advanced approach today is based on artificial neural networks (CASELI; NUNES, 2023). This latter approach, which benefits from the use of large datasets and deep neural networks, has become predominant in contemporary MT due to its ability to handle complex linguistic nuances.

2.2.1 Rule-Based Machine Translation

In the 1970s, Rule-Based MT dominated research in the field, dividing into three main approaches: direct translation (direct mapping between sentences), transfer-based translation (morphological and syntactic analysis), and interlingual translation (conversion into an abstract representation before the final translation) (GARG; AGARWAL, 2018).

As Hutchins e Somers (1992) explain, the Vauquois triangle (Figure 1) illustrates the hierarchical structure of linguistic processing involved in Rule-Based MT. At the bottom of the triangle lies the direct translation approach, which performs minimal linguistic analysis by mapping words or short phrases from the source language directly to their equivalents in the target language. In the middle level, the transfer approach introduces a deeper analysis by dividing the process into three stages: analysis of the source language, structural transfer between languages, and generation in the target language. This approach captures morphological and syntactic information, allowing better handling of structural differences.

Figure 1 – Vauquois triangle.



Source: Adapted from Hutchins e Somers (1992).

At the top of the triangle is the interlingual approach, which performs a complete linguistic analysis (morphological, syntactic, and semantic) of the source text in order to extract its underlying meaning. This meaning is then converted into a language-independent conceptual representation – often structured as a graph, tree, or symbolic logic – known as the interlingua, from which the target text is generated.

2.2.2 Statistical Machine Translation

The idea of automatic translation using statistical methods was initially proposed in 1949 by Warren Weaver, who suggested the application of mathematical techniques inspired by cryptography for language translation (BROWN et al., 1988). However, this approach gained traction only in the late 1980s and early 1990s, when IBM researchers revisited and developed practical statistical MT models, such as the IBM Word-Based Translation model, which represented a significant advancement in the field (HUTCHINS; SOMERS, 1992).

In addition to word-based Statistical MT, the phrase-based approach also had a significant impact and was widely successful in translating natural languages. This technique was extensively adopted and used by Google Translate from its launch in 2006 until 2016, when the tool began employing deep neural networks, improving both the quality and fluency of translations (KARAIVANOV; RAYCHEV; VECHEV, 2014).

❏ Word-based Statistical MT: aligns individual words between the source and target texts, calculating the probability of each translation correspondence. This method also allows for the deletion and insertion of words to better adjust the translation (CASELI; NUNES, 2023). The terms "source text" and "target text" refer, respectively, to the original text to be translated and the resulting text in the target language (NORD et al., 2016).

❏ Phrase-based Statistical MT: instead of isolated words, this method aligns blocks of words (n-grams) between the source and target texts. It compares them in the context of their neighboring phrases to enhance both fluency and translation accuracy (CASELI; NUNES, 2023).

2.2.3 Neural Machine Translation

The field of MT underwent a major paradigm shift. Statistical MT was largely replaced by Neural Machine Translation (NMT), which approaches translation using a single neural network (STAHLBERG, 2020). In NMT models, artificial neural networks are employed to transform an input sentence (source) into an output sentence (target).

This process typically involves an encoder–decoder architecture. The encoder reads the sentence in the source language and converts it into a compact vector representation that captures its semantic and contextual features. The decoder then uses this representation to generate the translation in the target language (GOODFELLOW; BENGIO; COURVILLE, 2016).

2.2.3.1 Encoder-Decoder with RNNs

One of the first Neural MT models was based on the encoder–decoder architecture using Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). An artificial neural network can be viewed as a structure composed of multiple interconnected processing units, called artificial neurons, organized in layers. Each neuron receives a numerical input and produces an output that is computed based on the weights and inputs associated with it. These values are combined and processed through a mathematical function that determines the output value, defining how information flows through the network to perform the translation task (CASELI; NUNES, 2023).

In this model, the encoder is an RNN that processes the input sequence and converts it into a compact vector representation that summarizes the contextual information of the sentence. This vector is then passed to the decoder – another RNN – which generates the output sequence in the target language (CASELI; NUNES, 2023).

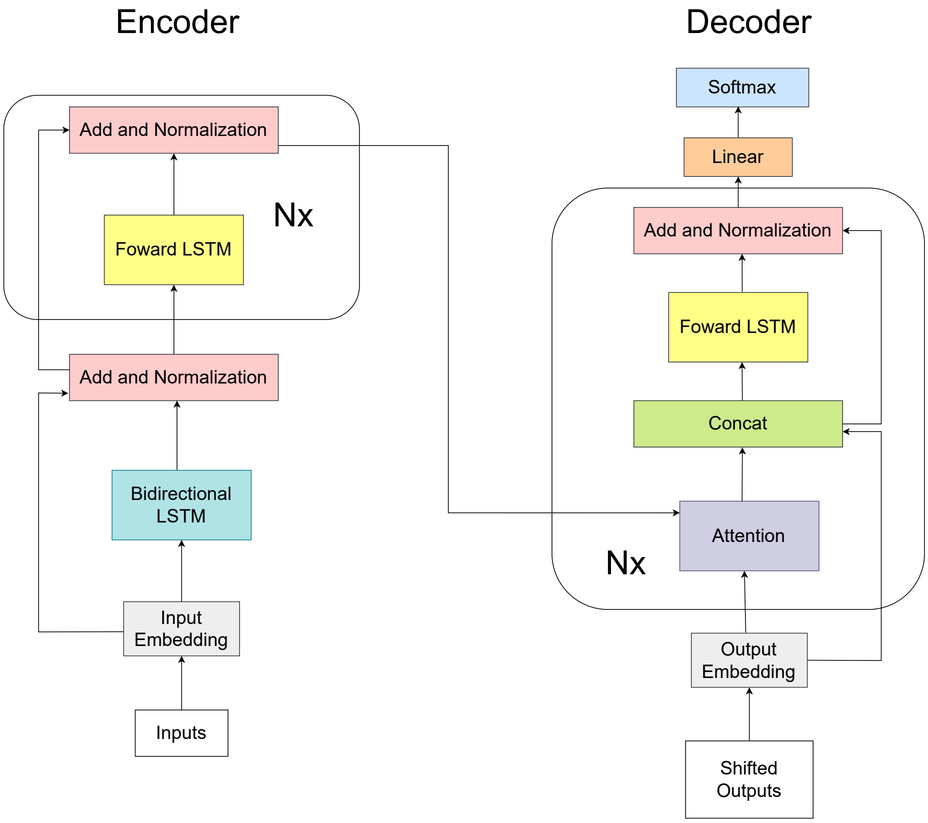
2.2.3.2 Encoder–Decoder with RNNs and Attention Mechanism

A significant evolution of the traditional encoder–decoder model was the introduction of the attention mechanism. This mechanism allows the model to “focus” on specific parts of the input sequence while generating each word in the translation, enabling it to better capture contextual information. With attention, the model does not need to rely solely on the last hidden states to produce the translation; instead, it can dynamically access information from different parts of the source sentence. This improvement enhances translation accuracy, especially for long and complex sentences, since the model can adjust its focus while generating each token in the target language (WU et al., 2016). Figure 2 illustrates the encoder–decoder architecture with RNNs and an attention mechanism used by Google between 2016 and 2020.

The architecture follows an encoder–decoder structure with an attention mechanism, allowing the decoder to selectively focus on relevant parts of the source sentence during translation. The encoder reads the input sentence (e.g., in French) and converts it into a set of context vectors that represent its meaning. It begins by embedding the input tokens into numerical vectors (input embeddings), which are then processed by a bidirectional LSTM that reads the sentence in both directions (left-to-right and right-to-left) to capture full contextual dependencies. The outputs are passed through multiple stacked LSTM layers with residual connections and normalization steps, allowing the network to be deeper and more stable during training. The final output of the encoder is a sequence of context vectors summarizing the entire input.

The decoder generates the translation in the target language (e.g., Portuguese), one word at a time. During training, the decoder is fed the correct previous word (shifted

Figure 2 – Model of the encoder–decoder architecture with RNNs and attention mechanism.



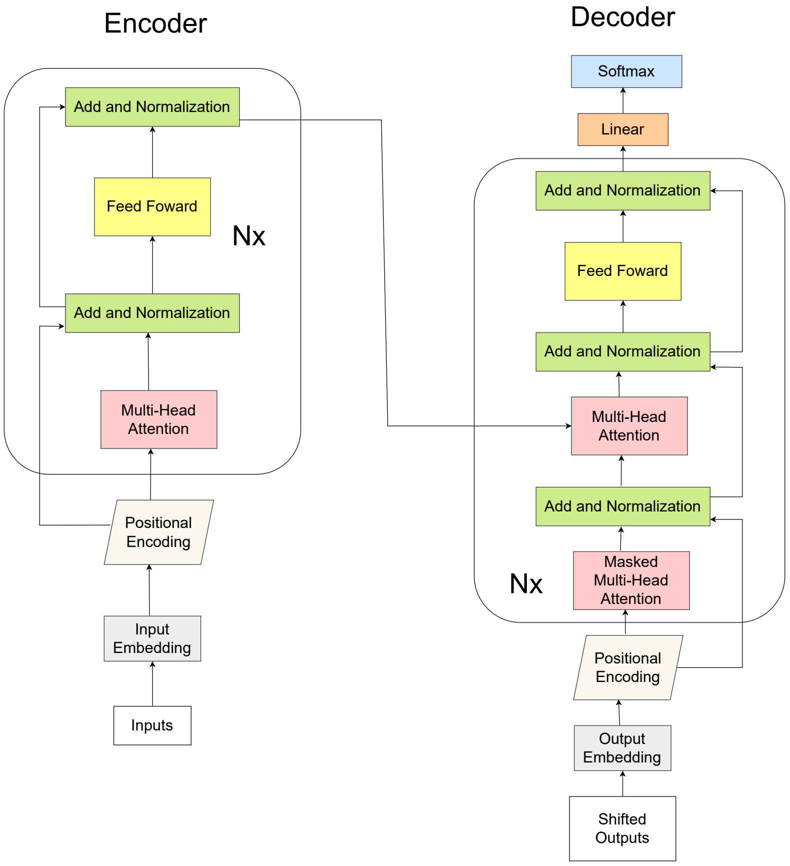
Source: Adapted from Stahlberg (2020).

output) and uses embeddings similar to the encoder. The attention mechanism computes alignment scores between the decoder's current state and all encoder outputs, producing a context vector that highlights the most relevant parts of the input for generating the next word. This vector is concatenated with the decoder's internal state and passed through stacked LSTM layers, followed by a linear and a softmax layer to predict the next token. Through this iterative process, the model dynamically adjusts its attention at each decoding step, generating fluent and contextually accurate translations.

2.2.3.3 Transformer

The Transformer architecture revolutionized the field of MT by completely replacing recurrent networks with attention layers. Unlike RNNs, which process words sequentially, the Transformer employs an attention mechanism that allows words to be processed simultaneously, capturing relationships between distant words in a sentence and providing a better understanding of the global context. This architecture consists of multiple layers of attention and feedforward neural networks that work efficiently, enabling training parallelization and reducing inference time (JURAFSKY; MARTIN, 2024). Figure 3 represents the Transformer architecture.

Figure 3 – Model of the Transformer architecture.



Source: Adapted from (VASWANI et al., 2017).

The Transformer model, introduced in the paper Vaswani et al. (2017), marked a shift from recurrent neural networks by relying exclusively on attention mechanisms. The encoder processes the input sentence (e.g., “the dog ran”) and transforms it into a set of contextual vectors. Each input token is first converted into a numerical vector through an embedding layer, followed by a positional encoding step that introduces information about word order – since the model itself does not process data sequentially. The encoder is composed of several identical layers, each containing two sublayers: a multi-head self-attention mechanism, which enables each word to attend to every other word in the sentence, and a feedforward neural network applied independently to each position. Residual connections and layer normalization are applied around each sublayer to ensure stable training in deep architectures.

The decoder generates the translated sentence one token at a time. It also uses multiple layers, each containing three subcomponents: masked multi-head self-attention, which prevents the model from accessing future words during training; multi-head attention, receiving inputs from both the encoder and decoder, which allows the decoder to focus on relevant parts of the input sentence; and a feedforward neural network. Finally, a linear transformation followed by a softmax layer predicts the most probable next word.

2.3 Specialized Machine Translation Models and LLMs

MT systems and Large Language Models (LLMs) represent two major paradigms in the evolution of automatic translation technologies. Specialized MT models, such as mBART, MarianMT, Google Translate, and OpenNMT with recurrent neural networks (RNNs), are typically based on encoder–decoder architectures trained on parallel corpora to establish direct correspondences between source and target sentences. These models emphasize linguistic alignment and rely on probabilistic or neural mechanisms to ensure consistency across language pairs. In contrast, LLMs, such as ChatGPT, and Maritaca AI, are founded on large-scale pretraining using extensive multilingual and multimodal data, enabling them to capture broader semantic and contextual relationships. This section provides an overview of the main MT models and LLMs discussed in this study.

mBART

The mBART (TANG et al., 2020) model (Multilingual Bidirectional and Auto-Regressive Transformer) is a sequence-to-sequence model pre-trained as a denoising autoencoder across multiple languages. It is based on the Transformer architecture, using an encoder–decoder structure where both components are trained jointly. mBART pre-trains both the encoder and decoder, enabling it to be directly fine-tuned for translation tasks. During pre-training, sentences from different languages are corrupted and reconstructed, forcing the model to learn a language-agnostic latent representation. When fine-tuned for MT, mBART can translate between any language pair present in its training corpus, including low-resource directions, benefiting from shared multilingual representations (TANG et al., 2020).

MarianMT

MarianMT (JUNCZYS-DOWMUNT et al., 2018) is part of the Marian Neural Machine Translation framework, a high-performance system written entirely in C++ for research and production use. It implements the Transformer-based encoder–decoder architecture and supports multi-GPU training and efficient deployment. MarianMT is particularly known for its integration with the Hugging Face Transformers library, which provides a large collection of pre-trained models for specific language pairs. The architecture follows the same principles as the Transformer, with self-attention layers in the encoder and cross-attention layers in the decoder; however, it is optimized for speed and scalability. This makes it suitable for large-scale multilingual translation scenarios (JUNCZYS-DOWMUNT et al., 2018).

OpenNMT with RNN

The OpenNMT (KLEIN et al., 2017) toolkit was one of the earliest open-source neural MT systems, originally implementing a recurrent neural network (RNN) encoder–decoder architecture with attention mechanisms. In this configuration, the encoder processes the input sentence sequentially, producing a context vector that summarizes the source information, while the decoder generates the output sequence word by word. The later versions of OpenNMT added support for Transformer. The attention mechanism allows the model to dynamically focus on different parts of the source sentence during decoding, improving long-sentence translations (KLEIN et al., 2017). Although later versions of OpenNMT introduced Transformer-based architectures, the RNN version remains available and continues to be used for research and educational purposes the rnn version is still available.

Google Translate

Google Translate [2] is a large-scale neural MT system developed by Google. Initially based on phrase-based statistical methods, it transitioned to neural MT in 2016, using deep LSTM networks with attention mechanisms to improve fluency and contextual understanding. GNMT introduced subword segmentation (Byte Pair Encoding) to handle rare words and achieved significant improvements in translation quality across multiple languages. More recent iterations have adopted Transformer-based architectures and multilingual pre-training, aligning with advances in large-scale multilingual modeling. The system continues to evolve with adaptive quality estimation and multilingual neural representations that allow zero-shot translation across unseen language pairs.

ChatGPT

ChatGPT[3], developed by OpenAI, is a large language model (LLM) designed for general-purpose text generation and comprehension. ChatGPT is built upon the GPT architecture (Generative Pre-trained Transformer), which relies solely on a Transformer decoder. ChatGPT can perform translation implicitly through its ability to model multilingual sequences learned from large-scale internet data. The model has been instruction-tuned using Reinforcement Learning from Human Feedback to follow user prompts more effectively.

[2] <https://research.google/blog/recent-advances-in-google-translate/>

[3] <https://openai.com/index/chatgpt/>

Maritaca AI

Maritaca AI (PIRES et al., 2023) is a Brazilian research initiative focused on developing open and culturally relevant large language models for Portuguese and multilingual applications. These models are based on the Transformer decoder-only architecture, similar to GPT, and are designed to perform various natural language processing tasks, including MT. Trained primarily on Portuguese corpora, Maritaca integrates data from diverse domains to ensure linguistic and contextual richness. Its training sources include educational materials such as essays and exams from the *Exame Nacional do Ensino Médio* (ENEM) – a nationwide standardized test in Brazil that assesses reading, writing, and reasoning skills – as well as legal texts in Portuguese. Due to its emphasis on Brazilian Portuguese and regionally grounded data, Maritaca is particularly valuable for evaluating translation quality involving Portuguese as a source or target language in diverse textual contexts (PIRES et al., 2023).

2.4 Evaluation Metrics

Automatic Translation Evaluation consists of examining translations generated by MT systems and assessing their quality based on predefined criteria. This evaluation can be performed either automatically or manually, with the analysis carried out by humans. However, it is important to note that "there is no single correct translation for a given source," as multiple correct translations may exist, making translation evaluation a complex problem (CASELI; NUNES, 2023).

2.4.1 Automatic Evaluation Metrics

Automatic evaluation metrics are widely used in MT research to provide quantitative assessments of translation quality. These metrics typically compare machine-generated translation with one or more human reference translations, measuring aspects such as lexical overlap, semantic similarity, and fluency.

2.4.1.1 BLEU

As human evaluations, despite offering higher quality assessments, are more expensive and time-consuming, IBM researchers proposed a “method for automatically evaluating MT that is fast, inexpensive, language-independent, and highly correlated with human judgment” (PAPINENI et al., 2002). This method is known as BLEU (Bilingual Evaluation Understudy).

Papineni et al. (2002) explain that the core idea of BLEU is that the closer an automatic translation is to a human reference, the better its quality. Implementing BLEU requires two main elements. The first is a numerical metric to measure the "closeness of the translation." The second is a corpus of high-quality human reference translations.

BLEU uses the concept of modified n-gram precision, a metric that counts matches between words or sequences of words (n-grams) in the candidate translation (the system output) and the reference translations (usually human-produced). The precision is adjusted to prevent words from being counted multiple times inappropriately, correcting over-generation.

Modified precision is computed by comparing the occurrences of n-grams in the candidate translation with their maximum occurrences in the reference translations. To avoid disproportionate penalties, the metric incorporates a brevity penalty, which accounts for cases where the candidate translation is shorter than the references. BLEU is calculated as the geometric mean of the n-gram precisions, weighted by an exponential factor that includes the brevity penalty. The BLEU score ranges from 0 to 1, with perfect scores being unlikely except when the candidate translation exactly matches the references. The general BLEU formula is expressed in Equation 1.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log(p_n)\right), \tag{1}$$

where:

1. BP is the brevity penalty that penalizes translations that are too short. Its formula is:

$$BP = \begin{cases} 1 & \text{if } c > r, \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases} \tag{2}$$

   with:

   ❑ c: total length of the candidate translation;

   ❑ r: total length of the closest reference translation.

   If $c > r$, no penalty is applied. Otherwise ($c \leq r$), BP introduces an exponential penalty based on the length difference.

2. p_n measures the precision of n-grams in the candidate translation, considering different n-gram sizes. Its formula is:

$$p_n = \frac{\sum_{\text{ngram}} \min(\text{Count\_cand}, \text{Count\_ref})}{\sum_{\text{ngram}} \text{Count\_cand}} \tag{3}$$

   ❑ Count_cand: number of times an n-gram appears in the candidate translation.

- ❏ Count_ref: maximum number of times the same n-gram appears in any of the reference translations.
- ❏ The min function prevents excessive repetitions from artificially inflating the score.

3. w_n are weights assigned to n-gram precisions. Usually, equal weights are assumed:

$$w_n = \frac{1}{N}, \quad \text{where } N \text{ is the maximum n-gram length considered.} \tag{4}$$

4. The logarithm is used to prevent longer n-grams from dominating the metric, while the exponentiation reverses the log effect, yielding the final BLEU score.

Despite its historical relevance and widespread adoption in machine translation research, BLEU has been increasingly criticized in recent years. As a surface-based metric relying on n-gram overlap, BLEU is limited in its ability to capture semantic equivalence, paraphrasing, and stylistic variation, which are common in high-quality translations (CALLISON-BURCH; OSBORNE; KOEHN, 2006). These limitations become even more pronounced in poetic translation, where translations can diverge lexically and syntactically from the reference while preserving meaning, rhythm, or aesthetic effect.

2.4.1.2 METEOR

Banerjee e Lavie (2005) present an alternative to BLEU. The authors explain that METEOR (*Metric for Evaluation of Translation with Explicit Ordering*) was developed to improve the assessment of MT quality, thereby overcoming some limitations of BLEU. One limitation of BLEU is that it does not give credit for approximate matches: even words that are semantically similar but not identical, such as the noun “responsibility” and the adjective “responsible,” are considered errors. METEOR, on the other hand, can identify these approximate matches through techniques such as *stemming*[4] and the use of synonyms or semantically related words (KOEHN, 2020).

The main approach of METEOR is to map words between the candidate translation and the reference translation. It compares translations based on *unigrams* (individual terms) and employs different types of mapping, such as exact match (for identical words), stem match, or synonym match. METEOR evaluates translation quality based on three main components: precision, recall, and fragmentation penalty. Unigram precision refers

[4] extracting the root of a word (CASELI; NUNES, 2023)

to the proportion of unigrams in the candidate translation that match the reference unigrams. Unigram recall is the proportion of unigrams in the reference that are found in the candidate translation. The fragmentation penalty measures deviations in word order between the candidate and the reference; the greater the disorder, the higher the penalty (BANERJEE; LAVIE, 2005). Equation 5 summarizes the relationship between these components.

$$\text{METEOR Score} = F_{\text{mean}} \times (1 - \text{Penalty}), \tag{5}$$

where:

1. F_{mean} is a variation of the harmonic mean that gives more weight to recall than to precision. The equation is:

$$F_{\text{mean}} = \frac{R + 9P}{10 \times P \times R}, \tag{6}$$

where:

- P = Precision: the proportion of words in the candidate that match the reference.
- R = Recall: the proportion of words in the reference that appear in the candidate.

2. *Penalty* penalizes deviations in word order using the formula:

$$\text{Penalty} = 0.5 \times \left(\frac{\#\textit{matched\_unigrams}}{\#\textit{chunks}} - 1 \right) \tag{7}$$

where:

- *chunks* refer to the number of contiguous sequences of words grouped based on unigram matches between the candidate and reference translations.
- *matched_unigrams* refer to the number of unigrams that match between the candidate and reference translations.

Example of Translations:

- **Candidate Translation:** “the president spoke to the audience”
- **Reference Translation:** “the president then spoke to the audience”

The first step is to identify unigrams that match between the two translations: "the", "president", "spoke", "to", "the", "audience". Next, adjacent unigrams in both translations are grouped, forming two *chunks*: "the president" and "spoke to the audience". Thus, the unigram matches result in two *chunks*.

How the Number of Chunks Affects the Penalty:

- **Fewer Chunks:** If unigrams match in larger n-grams (bigrams, trigrams, etc.), there are fewer chunks, indicating a more fluent and coherent translation alignment.
- **More Chunks:** If unigrams match individually, there are more chunks, indicating a less fluent and more fragmented correspondence.

Although METEOR represents a clear improvement over BLEU by incorporating recall, stemming, synonym matching, and word order penalties, it still presents important limitations for evaluating poetic translation. As a predominantly unigram-based metric, METEOR remains sensitive to surface lexical choices and local word alignments, which may not adequately reflect semantic adequacy or stylistic equivalence in creative texts. Despite achieving higher correlation with human judgments than BLEU in some settings, METEOR continues to struggle with paraphrastic variation and stylistic divergence, especially at the sentence level (CALLISON-BURCH; OSBORNE; KOEHN, 2006).

2.4.1.3 BERTScore

BERTScore is a metric for evaluating the quality of generated texts, such as translations, by comparing candidate sentences (generated by the model, $\hat{x}$) with references (human reference sentences, x). Unlike traditional metrics, such as BLEU, it does not rely on exact word matches, but instead uses contextual *embeddings* from the BERT model (based on the *Transformer*) to measure semantic similarity between *tokens* (ZHANG et al., 2020).

The first step of BERTScore, as explained by the authors, is to calculate the cosine similarity between the *tokens* of a candidate sentence ($\hat{x_j}$) and those of the reference (x_i), as in Equation 8. Since the *embeddings* provided by the BERT model are pre-normalized ($\|x_i\| = 1$), the formula simplifies as shown in Equation 9.

$$\text{sim}(x_i, \hat{x_j}) = \frac{x_i^\top \hat{x_j}}{\|x_i\|\|\hat{x_j}\|} \tag{8}$$

Where:

- $x_i^\top$ emphasizes that x_i has been transposed (turned into a row vector) before multiplying $\hat{x_j}$ (a column vector). The result $x_i^\top \hat{x_j}$ is the dot product, representing the sum of the products of the corresponding components of the two vectors.

$$\text{sim}(x_i, \hat{x_j}) = x_i^\top \hat{x}_j \tag{9}$$

After computing the similarity, BERTScore is obtained by combining precision and recall. Recall, Equation 10, compares each *token* of the reference (x) with all *tokens* of the candidate ($\hat{x}$), choosing the most similar *token*. The sum of maximum similarities is normalized by the number of reference *tokens* ($|x|$):

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x_j} \in \hat{x}} \text{sim}(x_i, \hat{x_j}) \tag{10}$$

Precision follows the same logic, comparing each *token* of the candidate ($\hat{x}$) with all *tokens* of the reference (x):

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{x_j \in \hat{x}} \max_{x_i \in x} \text{sim}(x_j, x_i) \tag{11}$$

From precision and recall, the F_1 score is calculated as in Equation 12:

$$F_{\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{12}$$

BERTScore can be adjusted using the Inverse Document Frequency (IDF) to give more weight to rare words. Equation 13 shows the IDF calculation:

$$\text{idf}(w) = \log \frac{1}{M} \sum_{i=1}^{M} I[w \in x^{(i)}] \tag{13}$$

Where:

❏ w: *token.*

❏ M: total number of reference sentences in the corpus.

❏ $I[w \in x^{(i)}]$: indicator function that returns 1 if w is present in sentence $x^{(i)}$, 0 otherwise.

The IDF-weighted recall and precision measures are given by:

$$R_{\text{BERT}} = \frac{\sum_{x_i \in x} \text{idf}(x_i) \max_{\hat{x_j} \in \hat{x}} \text{sim}(x_i, \hat{x_j})}{\sum_{x_i \in x} \text{idf}(x_i)} \tag{14}$$

$$P_{\text{BERT}} = \frac{\sum_{x_j \in \hat{x}} \text{idf}(x_j) \max_{x_i \in x} \text{sim}(x_j, x_i)}{\sum_{x_j \in \hat{x}} \text{idf}(x_j)} \tag{15}$$

(ZHANG et al., 2020) also explain Linear Rescaling (*Baseline Rescaling*) to make the results more interpretable:

$$\hat{R}_{\text{BERT}} = \frac{R_{\text{BERT}} - b}{1 - b} \tag{16}$$

Where b is the empirical baseline, calculated from random sentence pairs from large corpora, ensuring low lexical and semantic overlap. The same adjustment applies to P_{BERT} and F_{BERT}.

BERTScore represents a comparatively more appropriate alternative for evaluating poetic translation among commonly used automatic metrics. By relying on contextual embeddings rather than exact lexical overlap, BERTScore is better able to account for paraphrasing and semantic equivalence (ZHANG et al., 2020).

2.4.2 Human-Centered Evaluation of Translations

The most accurate evaluations use human evaluators. Human translations are typically assessed along two main dimensions: adequacy, which measures how well the translation preserves the exact meaning of the original sentence, also referred to as fidelity or accuracy, and fluency, which evaluates how natural, grammatical, and clear the translation is in the target language. Although human evaluation is more precise, automatic metrics are often used for practical reasons (JURAFSKY; MARTIN, 2024).

Jurafsky e Martin (2024) further explain that in fluency evaluation, evaluators analyze the translated text to determine how intelligible, clear, readable, and natural it is in the target language. A numerical scale is frequently used to rate sentences or paragraphs. In the adequacy dimension, if evaluators are bilingual, they can directly compare the source sentence with the proposed translation and assess how much of the original information has been preserved. If they are monolingual, the comparison is made between an automatic translation and a previously provided human reference, checking for semantic correspondence.

Another possible approach, according to the authors, is to present evaluators with pairs of candidate translations and ask them to choose the better one, allowing for a comparative evaluation without the need for numerical scores. This strategy is useful in scenarios where the goal is to identify preferences between different MT systems.

Although human evaluations are the most accurate, they are not always practical due to the costs and time involved. Manually assessing each translation requires significant resources; therefore, in many cases, automatic metrics are used. While not as precise as human evaluation, automatic assessments provide a faster and more accessible way to perform evaluations, especially when dealing with large volumes of translations (CHATZIKOUMI, 2020).

2.5 Topic Modeling

Topic modeling is a statistical and computational approach designed to uncover latent thematic structures within large text corpora (BLEI, 2012). It identifies recurring patterns of word co-occurrence, grouping terms that frequently appear together into distinct topics that represent underlying semantic themes. It can also reveal central motifs and diachronic thematic trends that might not be immediately perceptible through manual analysis (JELODAR et al., 2019).

2.5.1 BERTopic

BERTopic, presented by Grootendorst (2022), is a neural topic modeling framework that builds topic representations by combining transformer-based contextual embeddings with density-based clustering and a class-based TF–IDF ranking of terms. The method can be described step by step as follows.

1. **Contextual embedding:** Each document is encoded into a dense vector using a pretrained transformer encoder (for example, BERT or a multilingual variant).

2. **Dimensionality reduction:** Since transformer embeddings are high-dimensional, BERTopic reduces their dimensionality to a lower-dimensional manifold suitable for clustering. The framework commonly employs UMAP (Uniform Manifold Approximation and Projection) to preserve local and global structure while speeding up subsequent clustering.

3. **Clustering:** The reduced embeddings are clustered with HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). HDBSCAN automatically determines the number of clusters, identifies dense regions of the embedding space, and labels sparse points as noise (often assigned topic -1).

4. **Class-based Term Frequency-Inverse Document Frequency (c-TF-IDF):** For each cluster found by HDBSCAN, BERTopic aggregates all documents in the cluster into a single pseudo-document and computes a class-based TF–IDF to rank terms that are most representative of that cluster. A common formulation is:

$$\text{c-TFIDF}(t, C) = \frac{\mathrm{tf}_{t,C}}{\sum_{t'} \mathrm{tf}_{t',C}} \cdot \log\left(\frac{N}{1 + \mathrm{df}_t}\right),$$

   where $\mathrm{tf}_{t,C}$ is the frequency of term t in cluster C, df_t is the number of documents containing t in the whole corpus, and N is the total number of documents. Terms with the highest c-TF–IDF scores are selected as topic descriptors.

2.5.2 Topic Coherence

Topic modeling algorithms generate clusters of words representing latent themes. However, not all topics are equally meaningful. The C_V metric is widely used to assess the coherence of topics. It evaluates the degree to which the top words in a topic tend to co-occur within the corpus. The underlying assumption is that coherent topics contain words that frequently appear together in similar contexts (ROEDER; BOTH; HINNEBURG, 2015). Formally, the C_V score is computed as:

$$\mathrm{C\_V}(t) = \frac{1}{|P|} \sum_{(w_i,w_j)\in P} \mathrm{NPMI}(w_i, w_j), \tag{17}$$

where:

- t: the topic being evaluated.
- P: the set of all pairs of the top N representative words of the topic. For example, if the topic has words $\{w_1, w_2, w_3\}$, the pairs would be $(w_1, w_2), (w_1, w_3), (w_2, w_3)$.
- $\mathrm{NPMI}(w_i, w_j)$: normalized pointwise mutual information between words w_i and w_j, measuring how strongly the words co-occur relative to chance.
- $|P|$: the total number of word pairs in the topic, used to compute the average NPMI.

The NPMI is calculated as:

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log p(w_i, w_j)}, \tag{18}$$

where:

- $p(w_i, w_j)$: probability of co-occurrence of words w_i and w_j in the corpus.
- $p(w_i), p(w_j)$: probabilities of individual occurrences of w_i and w_j in the corpus.
- $\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}$: pointwise mutual information (PMI), which measures how much more frequently the words co-occur than expected if they were independent.
- $-\log p(w_i, w_j)$: normalization factor that scales PMI to a range between -1 and 1, producing the NPMI.

Higher C_V scores indicate more coherent topics, meaning that the words are semantically related and the topic is more interpretable.

2.6 Final Considerations

This chapter presented the theoretical foundations necessary for the development of this study. We discussed the notion of translation as a cultural and interpretive act and emphasized the specific challenges of poetic translation, where meaning, rhythm, sound, and imagery interact in ways that require balancing semantic fidelity and aesthetic effect. We then reviewed the historical trajectory of MT, from rule-based and statistical systems to neural models, highlighting the role of the Transformer architecture and comparing specialized MT systems (MarianMT, mBART, OpenNMT, and Google Translate) with large language models used for translation, such as ChatGPT and Maritaca AI.

We also examined the evaluation methods adopted in this work, including automatic metrics (BLEU, METEOR, BERTScore) and human-centered assessments that capture stylistic and poetic aspects often missed by purely computational measures. Finally, we presented topic modeling with BERTopic as a complementary analytical tool for examining thematic structures across translated poems. The next chapter discusses related work on poetic translation and previous applications of topic modeling, situating this research within other academic efforts in the field.

Chapter 3

Related Works

This chapter situates the present research within its academic context, reviewing studies that address the translation of literary and poetic texts, as well as research involving thematic analysis through topic modeling. Section 3.1, Machine Translation of Literary and Poetic Texts, examines previous work on the translation of poetry and other literary genres, highlighting approaches based on neural MT models and large language models, with attention to challenges related to style, rhythm, and meaning. Section 3.2, Topic Modeling, presents studies that apply topic modeling techniques, emphasizing recent advances such as BERTopic and their relevance for analyzing thematic patterns in poetry. Finally, Section 3.3 presents the final considerations of the chapter.

To conduct the systematic literature review, combinations of keywords were defined as search strings, involving terms related to poetry, translation, and automatic translation techniques, such as "poem translation", "poetry translation", and "literary text translation" combined with "machine learning", "LLM", "automatic", and "machine translation". Searches were carried out in well-established databases, including IEEE, Elsevier, and Google Scholar. The inclusion filters considered articles published from 2018 onward, focusing on works in the field of Computer Science or related to MT systems. The exclusion criteria removed articles that did not address Romance languages[1], did not include poetry in the study, or focused solely on automatic poetry generation without involving translation.

This systematic review revealed that most research in the field of MT has focused on translating technical texts, leaving the study of this technology's application to literary texts underexplored. This gap is particularly evident in poetry, especially in Romance languages (VáZQUEZ; MITKOV, 2023). Moreover, the review also indicated that there are few studies involving human evaluators for these languages, particularly for experts.

[1] Romance languages are a group of languages that developed from Vulgar Latin. Among the main Romance languages are French, Italian, Spanish, Portuguese, and Romanian. Source: <https://www.britannica.com/topic/Romance-languages>

3.1 Machine Translation of Literary and Poetic Texts

Many studies have explored the use of traditional artificial neural networks in translation across language pairs; however, only a limited number of works focus specifically on the translation of poetic texts. Humblé (2019) analyzed Google Translate's neural system in the translation of three English poems into Portuguese, showing that, despite its limitations, MT can be a valuable aid for post-editing. Vázquez e Mitkov (2023) investigated the influence of genre and historical period on the performance of neural translation systems, concluding that DeepL produced the most coherent and human-like translations. Focusing on form preservation, Ghazvininejad, Marjan, Choi, Yejin e Knight, Kevin (2018) developed neural models designed to maintain rhyme and rhythm in translations, with human evaluations confirming their effectiveness.

Song et al. (2023) explore Iterative Back-translation (IBT) to improve the quality of automatic poetry translation. The technique involves translating a text into the target language and then back-translating it into the original language, creating a parallel corpus that refines both stylistic and semantic adaptations. The study used a large set of poems in Chinese, English, and Portuguese, evaluating the translations with automatic metrics (BLEU, Bilingual Evaluation Understudy, and BERTScore) and human analysis. The results indicated that IBT outperformed the Baidu system in preserving stylistic features; however, it could introduce some semantic noise. The approach proved promising for enhancing poetic translation, with potential for expansion to multiple languages.

With the arrival of large language models (LLMs), the discussion among translators has shifted: instead of merely questioning the impact of these tools, the focus has turned to how to strategically use them. In this context, Resende e Hadley (2024) analyzed the capacity of LLMs in translating both rhymed and unrhymed poetry, comparing ChatGPT-3.5, ChatGPT-4, and Google Gemini with human translations. The study evaluated sonnets and free verses in Portuguese and Spanish translated into English, analyzing metrics such as lexical richness, lexical density, sentence length, and rhyme patterns. The results indicated that the models still face challenges in maintaining rhythmic structure and lexical diversity, particularly without specific instructions. The LLMs performed better with sonnets when guided by detailed instructions, but human translations showed greater stylistic fidelity.

In Chakrabarty, Saakyan e Muresan (2021), the authors explored mBART for neural poetry translation in a multilingual "many-to-one" (N-to-1) fine-tuning setup, translating poems from Russian, Spanish, Portuguese, German, Italian, and Dutch into English. They also evaluated the model in a "zero-shot" setting for Romanian, Ukrainian, and Swedish. mBART was trained on 190,000 poetic lines and was compared with non-poetic corpora. Human evaluations and automatic metrics (BLEU, BERTScore, and COMET) showed that multilingual fine-tuning on poetry outperformed bilingual approaches and fine-tuning on non-poetic texts. Table 1 summarizes the related works on MT of literature and poems.

Table 1 – Summary of related work on MT of literary and poetic texts.

Work	Languages	Model	Evaluation Metrics
Resende e Hadley (2024)	Portuguese to English; Spanish to English	ChatGPT-3.5, ChatGPT-4, Google Gemini	BLEU + human evaluation (2 experts)
Ghazvininejad, Marjan, Choi, Yejin e Knight, Kevin (2018)	French to English	Neural models developed by the authors: Model A: RNN-LSTM Model B: Biased Decoding (Unconstrained) Model C: Biased Decoding (All Potential Translation)	Mechanical Turk judges: A vs. B: 154 judges B vs. C: 84 judges Quality rating: 238 judges
Humblé (2019)	English to Portuguese	Google Translate	Human evaluation (1 expert)
Song et al. (2023)	English to Modern Chinese and vice-versa; English to Portuguese	Baidu Versions proposed by the authors: Back-Translation (BT) Iterative Back-Translation (IBT)	BLEU, BERTScore + human evaluation (4 experts)
Chakrabarty, Saakyan e Muresan (2021)	Russian, Spanish, Portuguese, German, Italian, Dutch to English Zero-shot: Romanian, Ukrainian, Swedish to English	mBART (multilingual fine-tuning)	BLEU, BERTScore, COMET + human evaluation (3 bilingual speakers per language)
Vázquez e Mitkov (2023)	English to Spanish	DeepL, Systran, Yandex	BLEU

Source: Own authorship.

The previous studies typically examine either LLMs or specialized neural MT models, this work directly compares both types of models in the translation of poetry. The evaluation examines six language pairs, including two Romance languages that remain underrepresented in poetic translation research (French and Portuguese) alongside English. Furthermore, this study combines automatic evaluation metrics and human judgment with topic modeling, allowing the analysis to address semantic content, poetic features, and thematic consistency. This study also incorporates fine-tuned models trained on song lyrics and poetic corpora, enabling a controlled comparison between general-purpose translation models and models adapted to literary style.

3.2 Topic Modeling

Topic modeling has increasingly been used to analyze thematic patterns in literature. Navarro-Colorado (2018) applied Latent Dirichlet Allocation (LDA), a model that identifies latent topics based on word co-occurrence, to a corpus of Spanish sonnets. The primary objective was to investigate what kinds of topics the model extracts from poetic language. The analysis demonstrated that the resulting topics often reflected stylistic and sound-based features rather than explicit semantic themes. Furthermore, the author argues that lemmatization may suppress poetic nuance and that entire poems – not isolated lines – should serve as semantic units.

More recent work has emphasized neural topic modeling. Babalola, Ojokoh e Boyinbode (2024) compared LDA, Non-negative Matrix Factorization (NMF), and BERTopic in the analysis of news headlines and demonstrated that BERTopic achieves greater semantic coherence, particularly in short and stylistically dense texts. BERTopic outperforms traditional models in capturing contextual meaning due to its use of transformer-based embeddings. These advantages stem from its use of contextual embeddings, making it particularly well-suited for short and stylistically complex texts such as poetry.

While BERTopic has been extensively applied in text mining tasks, its use in translation evaluation remains unexplored. Nevertheless, it offers promising potential as a complementary approach to compare thematic consistency across translations.

In contrast to previous studies, which primarily employ topic modeling to explore the thematic structures within a corpus, the present work adopts topic modeling as an evaluative component for poetic translation. Rather than using BERTopic solely to uncover themes, we apply it to compare thematic consistency among the original poems, the human reference translations, and their automatically translated versions across the language pairs. This approach allows us to analyze whether translation systems maintain not only semantic content but also the meaningful thematic associations present in the source text.

3.3 Final Considerations

The studies reviewed in this section highlight the importance of exploring approaches to literary MT, especially for poetry, while recognizing the nuances and complexities inherent in this type of text. Among the models analyzed, those based on the Transformer architecture have received the most attention in recent research, with a focus on systems like mBART and generative models such as ChatGPT. However, the literature lacks direct comparisons between these two categories (generative models and MT models), especially in poetic literary contexts. In light of these gaps, this study aims to compare the performance of both model types in the translation of poems across English, French, and Portuguese, using a multidimensional evaluation that combines automatic metrics, expert human analysis, and topic modeling.

The next chapter will detail the methodological framework adopted in this research. It includes the construction of the multilingual poetic corpus, the selection and configuration of the translation models, the procedures for automatic and human evaluation, and the application of BERTopic for evaluation using topic modeling.

Chapter 4

Proposed Approach for Comparing Automatic Poetry Translations

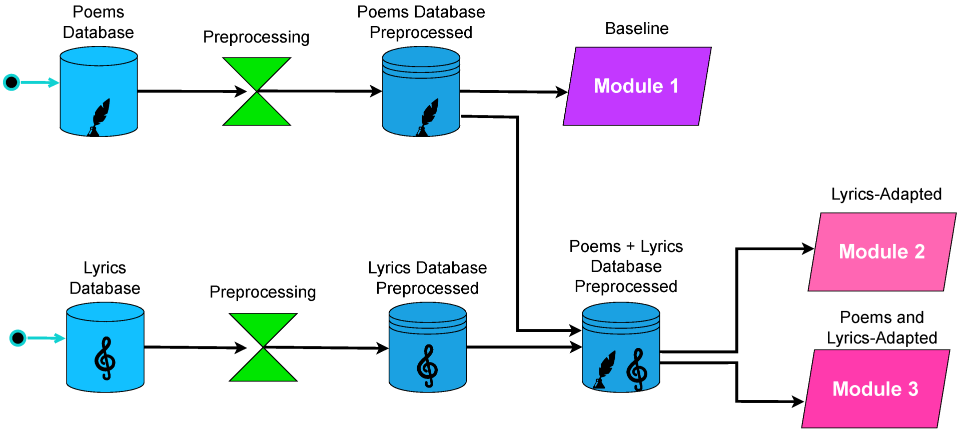
This chapter presents the methodology adopted for comparing automatic translations of poetry. Section 4.1 provides an overview of the proposed approach, outlining the main components of the workflow. Section 5.1 describes the datasets used in the experiments, including the corpus of poems and the corpus of song lyrics employed during fine-tuning. Section 4.2 details the preprocessing steps applied to the textual data prior to model training and evaluation. Section 4.3 introduces the translation models considered in this study and explains the fine-tuning strategies applied to those that support parameter adaptation. Section 4.4 presents the evaluation procedures, including automatic metrics, topic modeling analysis, and human expert assessment, which together form the basis for the results discussed in the following chapters. Finally, the last section presents the final considerations.

4.1 Overview

The main objective of this work is to evaluate the performance of automatic translation algorithms in the context of poetry, a task that requires not only semantic accuracy but also the preservation of stylistic and emotional elements. To achieve this, we propose a three-phase evaluation framework that integrates complementary perspectives: automatic metrics, thematic analysis, and human evaluation. In addition, we investigate fine-tuning strategies to adapt translation models more effectively to the specificities of poetic language. This comprehensive framework, summarized in Figure 4, combines quantitative and qualitative dimensions to compare automatic poetry translations.

The process begins with two databases: one of poems and another of song lyrics. Both corpora undergo a preprocessing stage to normalize and prepare the texts for subsequent experiments. After preprocessing, the database with poems is used in Module 1, while both the Poems and Lyrics Databases are used in Modules 2 and 3. From this stage

Figure 4 – Overall workflow of the proposed framework with the three modules analyzed.



Source: Own authorship.

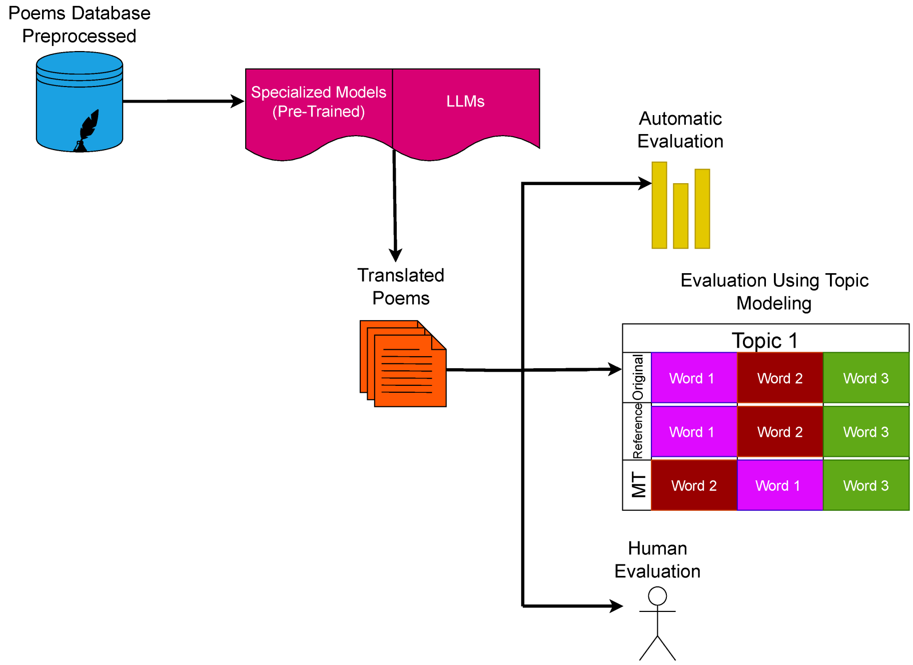
onward, the approach is organized into three modules of translation and evaluation.

Module 1 (Figure 5) establishes the baseline by using pre-trained translation models, serving as a reference for subsequent comparisons and embodying the core methodological proposal of this study: the three-phase evaluation framework. In this module, preprocessed poems are translated using both specialized MT models and LLM systems. The entire dataset is translated for each language pair, and the resulting translations are evaluated through three complementary approaches: automatic metrics, thematic analysis via topic modeling, and human evaluation. This module establishes the baseline for comparison with the subsequently fine-tuned approaches.

In addition to this initial framework, the approach explores strategies for improving translation performance through fine-tuning models that support it, such as MarianMT, mBART, and OpenNMT RNN, since fine-tuning is not feasible for closed systems like Google Translate. Two complementary strategies are explored: one based on song lyrics, which share linguistic and stylistic similarities with poetry, and another trained directly on poetry or on a combination of poetry and lyrics. Module 2 (Figure 6) focuses on the first strategy. Here, only the song lyrics from the Poems + Lyrics Database are initially used to fine-tune the specialized MT models. For the fine-tuning workflow, the song lyrics are separated and used to fine-tune the specialized MT models. For the testing workflow, the poems are used as input, and the models, now fine-tuned, translate the poems, with the results evaluated using automatic metrics.

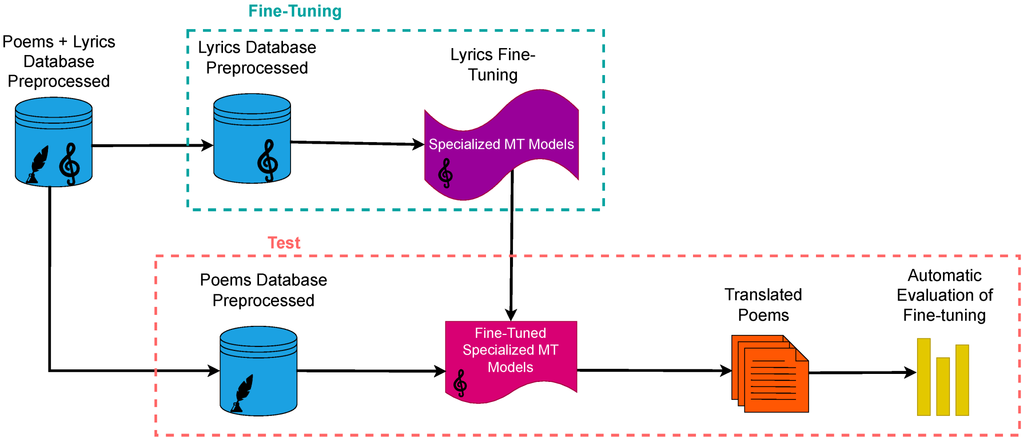
Module 3 (Figure 7) explores alternative fine-tuning strategies by comparing three approaches: models fine-tuned exclusively with poems, models fine-tuned with a combination of poems and song lyrics, and a baseline system without fine-tuning for comparison. The process is divided into two main phases: Fine-Tunings and Tests. In the Fine-Tunings phase, the first workflow involves fine-tuning specialized MT models using 80% of the preprocessed poems database. The second workflow follows a similar fine-tuning procedure

Figure 5 – Module 1 - Baseline: overview of the proposed methodology illustrating the three-phase evaluation framework for automatic poetry translations.



Source: Own authorship.

Figure 6 – Module 2 - Lyrics-Adapted: specialized MT models are fine-tuned using a lyrics corpus before translating poems.


Fine-Tuning
Poems + Lyrics Database Preprocessed
Lyrics Database Preprocessed
Lyrics Fine-Tuning
Specialized MT Models
Test
Poems Database Preprocessed
Fine-Tuned Specialized MT Models
Translated Poems
Automatic Evaluation of Fine-tuning

Source: Own authorship.

but combines the same 80% of poems with a preprocessed song lyrics dataset.

During the Tests, a single test dataset – corresponding to the remaining 20% of the poems – is employed to ensure comparability across all models. Three evaluation workflows are performed: (1) the first translates the test poems using MT models fine-tuned exclusively on poems; (2) the second uses models fine-tuned on both poems and lyrics; and (3) the third serves as the baseline, where the same set of poems is translated using pre-trained MT models and LLMs. Finally, the results from all three translation workflows are evaluated using automatic metrics.

4.2 Preprocessing

The following preprocessing steps were applied to all data (poems and lyrics): i) text cleaning using Python to remove non-linguistic symbols such as digits and special characters, while preserving punctuation, alphabetic characters with diacritics, whitespace, line breaks, and apostrophes, as these elements play a central role in poetic meaning, rhythm, and formal structure; ii) spacing normalization to eliminate extra spaces at the beginning and end of lines and to reduce multiple consecutive spaces between words; iii) manual verification of alignment between sentence pairs to ensure that the source column corresponded accurately to its respective translation in the target column; and iv) manual verification of quotation marks and commas to ensure proper balance and placement in the CSV structure. The example below illustrates the effect of the preprocessing steps on the text.

Before preprocessing:

> Let them repeat the link, and pour and pour&nbsp
> Their pleasure till they can no more!

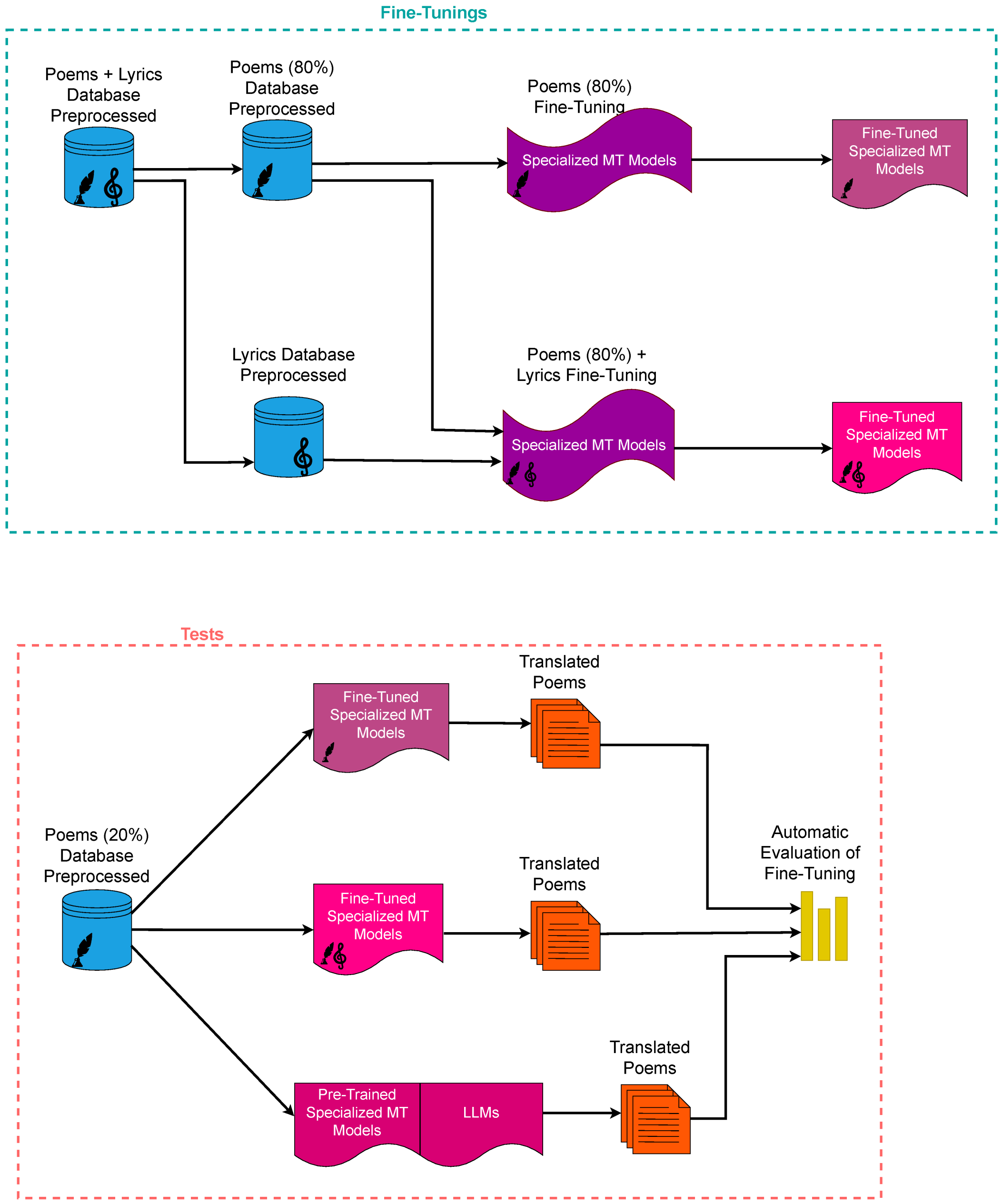
After preprocessing:

> Let them repeat the link, and pour and pour
> Their pleasure till they can no more!

Tokenization was handled differently for each model in order to leverage their native preprocessing mechanisms. The MarianMT and mBART models are based on *Transformers* (VASWANI et al., 2017) and have internal mechanisms to handle tokenization. However, tokenization must be explicitly called through the model's tokenizer function. On the other hand, closed systems like Google Translate, Maritaca AI, and ChatGPT accept raw natural language text directly, without the need for explicit tokenization.

On the other hand, for the OpenNMT (RNN) model, which relies on Recurrent Neural Networks, preprocessing must explicitly include both tokenization and vocabulary

Figure 7 – Module 3 - Poems and Lyrics-Adapted: comparative fine-tuning strategies, including poems-only fine-tuning, poems and lyrics fine-tuning, and baseline translation without fine-tuning.


Fine-Tunings
Poems + Lyrics Database Preprocessed
Poems (80%) Database Preprocessed
Poems (80%) Fine-Tuning
Specialized MT Models
Fine-Tuned Specialized MT Models
Lyrics Database Preprocessed
Poems (80%) + Lyrics Fine-Tuning
Specialized MT Models
Fine-Tuned Specialized MT Models
Tests
Fine-Tuned Specialized MT Models
Translated Poems
Poems (20%) Database Preprocessed
Automatic Evaluation of Fine-Tuning
Translated Poems
Fine-Tuned Specialized MT Models
Translated Poems
Pre-Trained Specialized MT Models
LLMs

Source: Own authorship.

creation. In this case, `sacremoses`[1] was employed for tokenization, and BPE[2] was used to build the vocabulary by segmenting tokens into subword units. This process enables the model to handle unknown or rare words more effectively during training.

4.3 Translation Models

The translation models selected for this study cover different approaches to MT and language modeling. First, we included Transformer-based models specialized in translation, such as MarianMT (JUNCZYS-DOWMUNT et al., 2018), mBART (TANG et al., 2020), and Google Translate[3]. To broaden the comparison, we also considered large language models (LLM), including ChatGPT-3.5[4], a general-purpose model, and Maritaca AI (PIRES et al., 2023), an LLM specifically trained for Brazilian Portuguese. In addition, we incorporated a classical approach based on Recurrent Neural Networks (RNNs), using OpenNMT (KLEIN et al., 2017), which contrasts with the Transformer-based architectures.

For the experiments, MarianMT, mBART, and OpenNMT (RNN) were tested both in their pre-trained form and after fine-tuning with the collected poems and song lyrics. After fine-tuning, the resulting models were saved and subsequently employed to translate the poetic texts. ChatGPT-3.5 and Maritaca AI were explored through two prompting strategies: one general and another specifically designed for poetry translation. Google Translate was employed in its default single-translation setting.

4.4 Evaluation

This study proposes three complementary evaluation strategies for assessing automatic poetry translations. The first strategy employs established automatic metrics – BLEU, METEOR and BERTScore – with the goal of quantitatively measuring lexical and semantic similarity between MT and human references. The second strategy applies topic modeling using BERTopic to evaluate thematic consistency across the original poems, human reference translations, and machine-generated outputs, aiming to identify whether the core themes are preserved or distorted during translation.

The third strategy involves expert human evaluation, conducted by a specialist, to assess translations based on qualitative criteria such as fluency, semantic preservation, poetic structure, and stylistic fidelity, covering aspects that automatic metrics are unable to capture, such as figurative language, rhyme, and rhythm. Together, these three ap-

[1] <https://pypi.org/project/sacremoses/>

[2] <https://huggingface.co/learn/nlp-course/chapter6/5>

[3] <https://research.google/blog/recent-advances-in-google-translate/>

[4] <https://openai.com/index/chatgpt/>

proaches provide a comprehensive framework for evaluating both the linguistic accuracy and the literary quality of automatic poetry translations.

4.4.1 Automatic Metrics

The BLEU metric (PAPINENI et al., 2002) evaluates translation quality by measuring the overlap of n-grams between the generated output and the reference, while also applying a brevity penalty to discourage overly short translations. In this work, it is computed using the `sentence_bleu` function from the `nltk.translate.bleu_score` module. METEOR (BANERJEE; LAVIE, 2005), in contrast, extends lexical comparison by incorporating stemming, synonymy matching, and penalties for differences in word order. The scores are obtained through the `meteor_score` function from `nltk.translate.meteor_score`, applied to tokenized inputs.

BERTScore (ZHANG et al., 2020) shifts the focus from surface-level similarity to semantic alignment, leveraging contextualized embeddings from BERT. By computing cosine similarity between these embeddings, it captures meaning beyond exact word matches and is therefore more robust to paraphrasing and lexical variation. In this study, we employ the `F1-score` function from the `bert_score` library to obtain the final evaluation scores.

4.4.2 Evaluation Using Topic Modeling

For the thematic analysis of the poetic corpus, BERTopic was selected due to its capacity to generate semantically coherent topics in short, stylistically rich texts such as poetry. BERTopic combines transformer-based sentence embeddings with clustering algorithms – UMAP for dimensionality reduction and HDBSCAN for density-based clustering – to identify topics by grouping semantically similar documents. To enhance topic representation, BERTopic applies c-TF-IDF (via the KeyBERT-inspired representation) to extract descriptive keywords per topic. (GROOTENDORST, 2022).

Embeddings were generated using the `distiluse-base-multilingual-cased-v2` model from the SentenceTransformers library, and a fixed seed was used to ensure reproducibility. To quantitatively assess the quality of the generated topics, the coherence score c_v was computed. The c_v metric combines the indirect cosine similarity of word embeddings with normalized pointwise mutual information (NPMI), offering a reliable measure of topic interpretability and correlation with human judgment (MANNING; SCHüTZE, 1999). The values of c_v range from 0 to 1, with higher values indicating better topic coherence.

In this study, BERTopic was applied exclusively to the original poetic texts, as the objective was to identify the thematic structure intended by the original authors. Since the translated versions, whether human or machine-generated, may introduce interpretative or stylistic variations that alter the thematic clustering, running separate topic models on each version would produce non-aligned and incomparable topic groups. Therefore, the topics identified in the original corpus were propagated to their corresponding translations, maintaining a consistent thematic reference across all versions. For each language pair, a list of representative words per topic was generated using class-based TF-IDF (c-TF-IDF), separately for the originals, the human reference translations, and the machine-generated outputs. When processing outputs from OpenNMT (RNN), the `<unk>` tokens were removed before extracting representative words, as these placeholders appear whenever the model fails to translate a word. These lists were then compared to analyze which topics were preserved and which diverged, allowing for an assessment of thematic consistency in the translation process.

4.4.3 Human Evaluation

For the human evaluation, the best translation outputs according to BERTScore were selected, following the same approach adopted by Chakrabarty et al. (CHAKRABARTY; SAAKYAN; MURESAN, 2021). The translation that achieved the highest BERTScore among the LLMs (ChatGPT 3.5 and Maritaca AI) and the highest among the specialized MT models (mBART, MarianMT, Google Translate, and OpenNMT with RNN) was selected and submitted to a human specialist for evaluation.

The evaluation was conducted using the original poem, the human reference translation, and the two selected automatic translations (one from an LLM and one from a specialized MT model), following criteria inspired by Gao et al. (GAO et al., 2024). The poetic aspects considered in the evaluation were defined based on the literary principles described by Goldstein (2005). The evaluation criteria are described below.

- ❑ **Poetic Structure:** Refers to the degree to which the translation preserves or recreates the structural characteristics of the poetry, while respecting the poetic conventions of the target language. The following aspects should be observed:
  - **Meter:** The regularity of the number of poetic syllables per line, respecting the rhythm of the original poem, or recreating an equivalent rhythm in the target language.
  - **Rhythm:** Sound and accentual patterns that produce musicality, including the alternation of stressed and unstressed syllables.
  - **Rhyme:** The preservation of the rhyme scheme or substitution with equivalent structures in the target language.
  - **Stanzas:** Maintenance of the structure in stanzas (tercets, quartets, sextains, etc.) and the number of lines per stanza.

  - **Enjambement:** When a line continues into the next without a pause, completing its meaning, one should observe whether this effect is recreated.

- ❑ **Poetic Stylistics:** Refers to the extent to which the translation preserves or recreates the stylistic characteristics of the poetry while respecting the poetic conventions of the target language. The following aspects should be observed if present in the original poem:

  - **Figures of Speech:** A non-literal way of using language to create emphasis, evoke emotion, or make comparisons.
  - **Stylistic Ambiguity:** Intentional ambiguities that contribute to the poem's polysemy and their preservation or recreation in the translation.
  - **Poetic Lexical Choice:** Use of elevated, archaic, colloquial, or otherwise marked vocabulary, consistent with the style of the original poem.

- ❑ **Fluency:** Evaluates the naturalness of the translated text in the target language, considering whether the lexical choices and syntax align with the poetic style in that language.

- ❑ **Meaning Preservation:** Refers to the ability of the translation to convey the meaning of the original text.

- ❑ **Overall Impression:** A general evaluation of translation quality that considers all previous aspects. The central question is whether the poem functions in the target language as an autonomous work, preserving the essence of the original in a harmonious manner.

To make the evaluation numerical and facilitate comparison between translations, we adopted a scale from 1 to 5, where 1 indicates that the translation fails to meet the criteria and 5 indicates that it fully satisfies them.

4.5 Final Considerations

This chapter presented the methodological framework proposed in this study for evaluating automatic poetry translation. The approach integrates a framework with three complementary dimensions of analysis: (i) automatic evaluation metrics, which provide quantitative measures of lexical and semantic similarity; (ii) thematic analysis through topic modeling, which examines whether the thematic structure of poems is preserved across translations; and (iii) expert human evaluation, which assesses qualitative aspects such as poetic structure, stylistic fidelity, and overall aesthetic impact. Together, these three phases form a comprehensive evaluation pipeline designed to capture both linguistic accuracy and poetic expressiveness.

The chapter also introduced a modular workflow for translation and fine-tuning. Module 1 establishes the baseline by translating poems using pre-trained MT models and LLMs, and evaluating their outputs using the three-phase framework. Module 2 explores whether fine-tuning with song lyrics can improve translation performance. Module 3 expands this investigation by comparing models fine-tuned exclusively on poems, models fine-tuned on a combined corpus of poems and lyrics, and non-fine-tuned baselines. This modular structure allows a systematic comparison of how different forms of domain adaptation influence translation quality, particularly in the context of poetic texts. In the next chapter, we present the experimental results obtained from applying this methodology.

Chapter 5

Experiments and Analysis of Results for Comparing Automatic Poetry Translations

This chapter presents and analyzes the experimental results obtained from the proposed framework for evaluating automatic poetry translation. Section 5.2 first describes the parameter settings and implementation details adopted for each model. Section 5.3 reports the results of Module 1, which establishes the baseline comparison across models using a three-phase evaluation pipeline: automatic metrics, evaluation with topic modeling, and human evaluation. Section 5.4 then presents the results of Module 2, where specialized MT models are fine-tuned using song lyrics. Section 5.5 discusses the results of Module 3, comparing models fine-tuned exclusively on poems, models fine-tuned on a hybrid poems+lyrics corpus, and non-fine-tuned baseline systems. The last section presents the final considerations of the chapter.

5.1 Datasets

We collected a total of 300 poems and their respective translations in six language pairs, amounting to 1,800 poems and 1,800 translations. The language pairs are: French-English (FR-EN), French-Portuguese (FR-PT), English-French (EN-FR), English-Portuguese (EN-PT), Portuguese-French (PT-FR), and Portuguese-English (PT-EN). The poems were sourced from various websites[1] with human translations obtained through a combination of web scraping and manual collection, depending on the website. All websites consulted for the dataset (BORGES, 2025), along with the source code, are publicly available in the project repository. The data were collected between January 2, 2025, and February 12, 2025.

[1] All the poems are either in the public domain or available for use for research and study purposes.

The dataset was employed both for translating the complete set of 300 poems using each pre-trained model and for the fine-tuning modules. For fine-tuning, a holdout strategy was applied to partition the data, following the experiments proposed by Bichri, Chergui, and Hain (BICHRI; CHERGUI; HAIN, 2024). Also, the holdout approach was chosen because it allows explicit control over data separation, making it possible to ensure that poems by different poets were distributed across the splits rather than concentrated in a single subset. Specifically, 80% of the dataset was randomly allocated to the combined training and validation sets (with 60% used for training and 20% for validation), while the remaining 20% was reserved for testing.

In addition to the poem corpus, we compiled a parallel dataset of song lyrics for the fine-tuning stage. This dataset comprises bilingual lyrics across the same six language pairs, with the following distribution: English-French (8,308), English-Portuguese (8,106), French-English (9,675), French-Portuguese (3,411), Portuguese-English (9,795), and Portuguese-French (9,390), totaling 48,685 aligned lyric pairs. All lyric pairs were used during the training phase of the fine-tuning process. Table 2 provides a summary of the datasets employed in this study.

Table 2 – Summary of the datasets employed.

Language Pair	**Poems**	**Lyrics**
French–Portuguese	300	3,411
French–English	300	9,675
English–Portuguese	300	8,106
English–French	300	8,308
Portuguese–French	300	9,390
Portuguese–English	300	9,795
Total	1,800	48,685

Source: Own authorship.

5.2 Parameterization of the Models

In order to ensure comparability across systems, each translation model was parameterized according to either the configurations suggested in related works or the official documentation of the respective frameworks. For models that supported fine-tuning, the validation set was used to monitor training performance and implement early stopping, thereby preventing overfitting. All experiments were conducted on an NVIDIA TITAN Xp GPU with 12GB of memory, running CUDA version 12.4. Below, we present the specific parameterization adopted for each model used in this study.

- **mBART**: The model `facebook/mbart-large-50-many-to-many-mmt` was used with its tokenizer. Translations were generated by explicitly forcing the output language using the `forced_bos_token_id` parameter, due to the model's inconsistent detection of Portuguese. Fine-tuning followed the parameters from Chakrabarty, Saakyan e Muresan (2021), with 3 epochs, learning rate of 2e-5, batch size of 8 and gradient accumulation of 10.

❑ **MarianMT**: Used the Helsinki-NLP models[2] for pre-training. For unsupported pairs (e.g., French-Portuguese), a two-step translation via English was required. Tokenization was performed using `MarianTokenizer`. Fine-tuning followed the same configuration as mBART.

❑ **Google Translate**: Automatic translations were also generated using Google Translate[3].

❑ **OpenNMT (RNN)**: Initially trained with the TED2020 corpus[4] using an LSTM model (1 layer, hidden size 512, batch 16). The training configurations were based on the settings provided in the OpenNMT repository[5]. After 100,000 steps, translations were generated. Fine-tuning was done using a reduced `train_steps` of 10,000. Tokenization, segmentation, and vocabulary generation were done via OpenNMT tools.

❑ **ChatGPT 3.5 and Maritaca AI**: The model used for ChatGPT translations was ChatGPT 3.5 Turbo, while for Maritaca AI, the Sabiazinho-3 model was selected. The API implementation in Python followed the official instructions provided on the OpenAI[6] and Maritaca AI[7] platforms. Each poem in the dataset for every language pair was translated using two different prompts, designed to explore the influence of prompt specificity on translation quality and to compare the model's performance under more general versus more targeted instructions. The two prompts used are presented below.

  – "Translate the following poem from source to target", a more general instruction;

  – "Translate the following poem from source to target, maintaining the poetic style and rhymes", a more specific one.

The following tables summarize the parameterization of the translation models used. Table 3 shows the configurations for fine-tuned models (mBART, MarianMT, and OpenNMT), including key hyperparameters and tokenization details. Table 4 presents the

[2] <https://huggingface.co/Helsinki-NLP>

[3] <https://translate.google.com/>

[4] <https://opus.nlpl.eu/>

[5] <https://github.com/OpenNMT/OpenNMT-py/blob/master/config/config-rnn-summarization.yml>

[6] <https://platform.openai.com/docs/guides/text?api-mode=chat>

[7] <https://docs.maritaca.ai/pt/visao-geral>

closed systems (Google Translate, ChatGPT, and Maritaca AI), indicating model identification and prompt settings.

Table 3 – Parameterization of specialized translation models with fine-tuning.

Parameter	**mBART**	**MarianMT**	**OpenNMT (RNN)**
Model	mbart-large-50-many-to-many-mmt	Helsinki-NLP	LSTM (1 layer, hidden 512)
Pre-training / Base	mBART-large-50	Helsinki-NLP	TED2020 corpus
Epoch	3	3	-
Learning Rate	2e-5	2e-5	-
Batch Size	8	8	16
Gradient Accumulation	10	10	-
Steps / Train Steps	-	-	100,000 pre-train / 10,000 fine-tune
Tokenizer	mBART tokenizer	MarianTokenizer	OpenNMT tokenizer

Source: Own authorship.

Table 4 – Parameterization of closed translation systems.

Parameter	**Google Translate**	**ChatGPT**	**Maritaca AI**
Model	-	ChatGPT 3.5 Turbo	Sabiazinho-3
Prompt Details	-	Two prompts: general and poetic-style preserving	Two prompts: general and poetic-style preserving

Source: Own authorship.

5.3 Results of Module 1 - Baseline

Module 1 establishes the baseline by translating all poems in the dataset with pre-trained MT models and LLMs. In this module, all poems are used for model evaluation, as the models are pre-trained and there is no training phase for fine-tuning. The resulting translations are assessed using automatic metrics, thematic analysis, and human evaluation.

5.3.1 Automatic Evaluation

The following analysis reports the quantitative results of the baseline models, covering both the specialized pre-trained MT models (Table 5) and the LLMs (Table 6) with general and specific prompts, evaluated through automatic metrics such as BLEU, METEOR and BERTScore. These results provide an initial comparative perspective on how the selected models perform when directly applied to poetry without additional adaptation. Bold values indicate the best result among the models for each metric in each language pair.

Table 5 – Automatic evaluation metrics for poetry translations across specialized MT models.

Metric	Language Pair	mBART	MarianMT	Google Translate	OpenNMT (RNN)
BLEU	FR-EN	0.2751	0.2948	**0.3920**	0.0526
	FR-PT	0.0628	0.1650	**0.1973**	0.0303
	EN-FR	0.1421	0.1805	**0.2528**	0.0226
	EN-PT	0.1163	0.1480	**0.1821**	0.0143
	PT-FR	0.0817	0.1899	**0.2389**	0.0586
	PT-EN	0.1403	0.2411	**0.3472**	0.0546
METEOR	FR-EN	0.5462	0.5755	**0.6560**	0.2632
	FR-PT	0.2339	0.4241	**0.4653**	0.2062
	EN-FR	0.3886	0.4369	**0.5124**	0.1911
	EN-PT	0.3781	0.4202	**0.4583**	0.1934
	PT-FR	0.2562	0.4049	**0.5405**	0.2337
	PT-EN	0.4202	0.5213	**0.6152**	0.2991
BERTScore	FR-EN	0.9029	0.9024	**0.9202**	0.7977
	FR-PT	0.7051	0.7832	**0.8125**	0.6013
	EN-FR	0.7812	0.7989	**0.8276**	0.5759
	EN-PT	0.7737	0.7822	**0.8082**	0.5762
	PT-FR	0.7045	0.7657	**0.8031**	0.6383
	PT-EN	0.8628	0.8980	**0.9229**	0.8049

Source: Own authorship.

Table 6 – Automatic evaluation metrics for poetry translations across LLMs.

Metric	Language Pair	ChatGPT (Prompt1)	ChatGPT (Prompt2)	Maritaca (Prompt1)	Maritaca (Prompt2)
BLEU	FR-EN	0.3828	0.3527	0.3590	**0.3890**
	FR-PT	0.2490	0.2446	0.2476	**0.3337**
	EN-FR	**0.2411**	0.2195	0.2094	0.2011
	EN-PT	0.1807	0.1786	0.1750	**0.2251**
	PT-FR	**0.2716**	0.2639	0.2491	0.2472
	PT-EN	**0.3310**	0.3074	0.3075	0.3209
METEOR	FR-EN	0.6435	0.6195	0.6307	**0.6504**
	FR-PT	0.5110	0.5082	0.5165	**0.5776**
	EN-FR	**0.5001**	0.4753	0.4733	0.4515
	EN-PT	0.4601	0.4532	0.4525	**0.4863**
	PT-FR	**0.4802**	0.4715	0.4608	0.4505
	PT-EN	**0.6004**	0.5821	0.5823	0.5855
BERTScore	FR-EN	**0.9263**	0.9209	0.9201	0.9191
	FR-PT	0.8210	0.8197	0.8170	**0.8322**
	EN-FR	**0.8211**	0.8076	0.8167	0.8026
	EN-PT	0.8006	0.7970	0.8011	**0.8117**
	PT-FR	**0.8093**	0.8017	0.8003	0.7926
	PT-EN	**0.9238**	0.9187	0.9182	0.9195

Source: Own authorship.

5.3.1.1 Model Performance Analysis

Analyzing the specialized MT systems, Google Translate achieved the best overall performance, with the highest scores in nearly all metrics and language pairs. For example, in the PT-EN direction, it scored 34.72% in BLEU, 61.52% in METEOR, and 89.80% in BERTScore. This performance reflects Google's access to massive multilingual corpora and continuous model updates, which likely help maintain high translation quality across both high- and low-resource languages.

MarianMT, also based on the Transformer architecture, consistently performed below Google Translate but above mBART and OpenNMT. Its targeted design for high-quality parallel corpora enables relatively robust translations, particularly for PT-EN and FR-EN. For instance, MarianMT achieved 24.11% BLEU, 52.13% METEOR, and 89.80% BERTScore for PT-EN, outperforming mBART in all metrics.

mBART, as a multilingual sequence-to-sequence model, showed intermediate performance across most language pairs. While it benefits from broad multilingual capabilities, this generalist training sometimes leads to suboptimal results. For example, mBART scored 14.03% BLEU, 42.02% METEOR, and 86.28% BERTScore in PT-EN, clearly below MarianMT and Google Translate.

OpenNMT, based solely on an RNN architecture, presented the lowest results among the evaluated models, highlighting the importance of the shift toward transformer-based architectures. In the EN-PT direction, BERTScore was only 57.62% confirming the lowest performance among all models. This underscores that RNN-only approaches struggle to capture long-range dependencies and complex linguistic structures – crucial elements in poetic translation.

For the large language models (LLMs), ChatGPT and Maritaca generally outperformed the specialized MT systems across most metrics and language pairs. ChatGPT with Prompt 1, a generalist and direct prompt, typically produced the highest BLEU and METEOR scores across multiple directions, such as 33.10% BLEU, 60.04% METEOR and 92.38% BERTScore for PT-EN. Maritaca Prompt 2 often matched or exceeded Prompt 1 in semantic and stylistic metrics. For example, in FR-PT, Maritaca Prompt 2 achieved 83.22% BERTScore, surpassing Prompt 1's 81.70%.

Google Translate stood out among the specialized MT models, especially when evaluated with surface-oriented metrics such as BLEU and METEOR. When considering BERTScore, which relies on contextual embeddings and is therefore more suitable for assessing semantic similarity in poetry, LLMs obtained higher scores in most language pairs. Nevertheless, Google Translate still produced highly competitive results according to the automatic metrics.

5.3.1.2 Prompt Analysis

The choice of prompt had a significant impact on the performance of the LLMs. In the case of ChatGPT, Prompt 1 generally outperformed Prompt 2 in BLEU and METEOR across most language pairs. For instance, in the PT-EN pair, BLEU with Prompt 1 was 33.10%, compared to 30.74% with Prompt 2. Similarly, METEOR was 60.04% with Prompt 1 versus 58.21% with Prompt 2. Although BERTScore values were relatively close between the prompts, Prompt 1 also achieved slightly higher scores in most directions, such as FR-EN (85.06% vs. 84.14%) and PT-FR (80.93% vs. 80.17%). A representative example from the EN-PT pair illustrates this trend, where BERTScore was 80.06% with Prompt 1 versus 79.70% with Prompt 2. As shown in Table 7, Prompt 2 alters the second verse from "it is shadowed green" to "sombrias são suas cores," introducing stylistic ornamentation at the expense of semantic fidelity.

Overall, ChatGPT Prompt 1, which was more general and direct, typically produced translations that were closer in meaning to the original text, resulting in higher metric scores in most language pairs. In contrast, Prompt 2 explicitly asked for poetic elements such as rhyme, rhythm, or stylistic embellishments. While this approach sometimes yielded more creative or aesthetically elaborate outputs, it often introduced semantic deviations, which negatively affected automatic evaluation metrics that prioritize fidelity to the reference.

Table 7 – Example illustrating the differences of Prompt 1 and Prompt 2 on ChatGPT's translation (EN-PT).

Original	**Reference**
land lies in water; it is shadowed green.	terra entre águas, sombreada de verde.
Prompt 1	**Prompt 2**
A terra repousa na água; é sombreada de verde.	A terra repousa na água; sombrias são suas cores.

Source: Own authorship.

Conversely, Maritaca AI demonstrated greater consistency and balance between prompts. While Prompt 1 still performed slightly better in a few scenarios, Prompt 2 demonstrated good outcomes. In the EN-PT pair, Prompt 2 achieved a BERTScore of 81.17%, surpassing the 80.11% obtained by Prompt 1. In BLEU, the gains were more modest but followed a similar trend in some cases, such as FR-EN (38.90% with Prompt 2 vs. 35.90% with Prompt 1). In METEOR, several improvements were also observed, such as in FR-PT (57.76% with Prompt 2 vs. 51.65% with Prompt 1). These results suggest that Maritaca was able to integrate poetic instructions with greater fluency while maintaining semantic alignment.

A particularly illustrative case appears in the EN-PT direction with the translation of the verse "it is shadowed green." Prompt 1 produced "terra jaz na água; é sombreada de verde," whereas Prompt 2 yielded "Terra entre águas, sombreada de verde," aligning closely with the human reference. As a result, Prompt 2 achieved a higher BERTScore of 81.17%, compared to 80.11% with Prompt 1. The poetic enhancements introduced by Prompt 2 did not compromise meaning and even improved rhythmic and syntactic alignment. This suggests that Maritaca AI was able to benefit from stylistic prompts without introducing distortions.

Another observation with the Maritaca prompt 2 is that translations into Portuguese achieved the highest scores across all metrics when using this prompt. For example,

Table 8 – Example illustrating the differences of Prompt 1 and Prompt 2 on Maritaca's translation (EN-PT).

Original	**Reference**
land lies in water; it is shadowed green.	terra entre águas, sombreada de verde.
Prompt 1	**Prompt 2**
terra jaz na água; é sombreada de verde.	Terra entre águas, sombreada de verde.

Source: Own authorship.

BERTScore reached 83.22% for FR-PT and 81.17% for EN-PT, with BLEU and METEOR showing the same trend. This outcome is consistent with expectations, as Maritaca was primarily trained on Portuguese data.

5.3.1.3 Evaluation Metrics Analysis

Among the applied metrics, BLEU presented the lowest scores. For example, in the FR-EN pair, even the best-performing specialized MT model, Google Translate, achieved only 39.20%, while most other models remained below 30%. A similar trend is observed among the LLMs, with BLEU scores ranging between 20% and 30%, whereas the other metrics generally exceeded 60%. This outcome is expected, since BLEU relies exclusively on exact n-gram overlap between the generated and reference translations, penalizing lexical or structural variations that may still be correct but deviate from the literal form of the text. METEOR, which accounts for synonymy, stemming, and partial matches, generally produced scores between BLEU and BERTScore, though still substantially lower than BERTScore.

BERTScore consistently produced the highest scores across all models and language pairs. This metric evaluates semantic similarity using contextual embeddings rather than strict n-gram overlap, allowing it to recognize translations that convey the correct meaning, even if the phrasing differs from the reference. For instance, in the FR-EN pair, Google Translate achieved a BERTScore of 92.02%, and MarianMT scored 90.24%, which is substantially higher than the corresponding BLEU scores. BERTScore's robustness to lexical and syntactic variations makes it particularly suitable for evaluating poetic translation, where faithful meaning is more important than a literal word-for-word reproduction. Although alternative pretrained embedding models could be used within BERTScore, as an alternative to the `bert-base-multilingual-cased` model adopted, such variations, for example, `roberta-large` for english, did not change the ranking of the best-performing models and only minimally affected the reported percentages.

5.3.1.4 Language Pair Performance

The analysis of language pair performance across metrics reveals a consistent trend: translations into English generally obtained the highest scores across all models. For the specialized MT systems, FR-EN and PT-EN pairs reached the best BLEU, METEOR, and BERTScore values in Google Translate (39.20% and 34.72% BLEU; 65.60% and 61.52% METEOR; 92.02% and 92.29% BERTScore, respectively). MarianMT and mBart followed a similar pattern, with PT-EN and FR-EN emerging as their strongest directions. These outcomes align with expectations, as English benefits from being a high-resource language, with abundant corpora, dictionaries, annotated datasets, and NLP tools (KHAN et al., 2023), whereas low-resource languages suffer from a stark shortage of such resources.

In contrast, translations into Portuguese tend to produce lower results overall. FR-PT often stands out as the weakest direction across metrics, particularly for mBART (BLEU: 6.28%; METEOR: 23.39%; BERTScore: 70.51%). EN-PT also frequently presents low scores, showing that translating into Portuguese remains a challenging task. This is especially evident in Google Translate, where EN-PT achieved the lowest results across all metrics.

For the LLMs, the trend is similar: translations into English generally achieved the highest scores across models and prompts. Both ChatGPT prompts 1 and 2 and Maritaca prompt 1 obtained their best BLEU, METEOR, and BERTScore results for FR-EN, while Maritaca prompt 2 reached its top scores for PT-EN. For example, BERTScore for FR-EN was 92.63% in ChatGPT prompt 1, 92.09% in ChatGPT prompt 2, and 92.01% in Maritaca prompt 1, with PT-EN in Maritaca prompt 2 reaching 91.95%. BLEU and METEOR follow the same pattern, confirming English as the dominant target language.

Similarly to the specialized MT models, EN-PT also shows low performance in LLMs. In ChatGPT, BERTScore reaches its lowest values for EN-PT, at 80.06% and 79.70% in Prompts 1 and 2, respectively. For Maritaca, the lowest BERTScore values are observed for PT-FR, with 80.03% and 79.26% in Prompts 1 and 2. These trends reinforce the general challenge of translating into low-resource languages, where limited training data constrain model performance.

5.3.1.5 Statistical Analysis

To statistically assess whether the differences in translation quality among the models are significant, we applied the Friedman test, a non-parametric statistical test commonly used to detect differences in performance across multiple models over multiple datasets or tasks. In our case, the test was conducted for each language pair and each evaluation metric (BLEU, METEOR, and BERTScore), comparing all translation systems used in the study.

As shown in Table 9, the Friedman test results indicate statistically significant differences across the models for all language pairs and all metrics, with extremely low p-values (e.g., $p < 1e\text{-}228$ for BLEU in the FR-EN pair), much smaller than the conventional threshold of 0.05. These results provide strong evidence to reject the null hypothesis that all models perform equally. In other words, the translation models do not perform equally, and at least one model performs significantly better or worse than the others in each comparison.

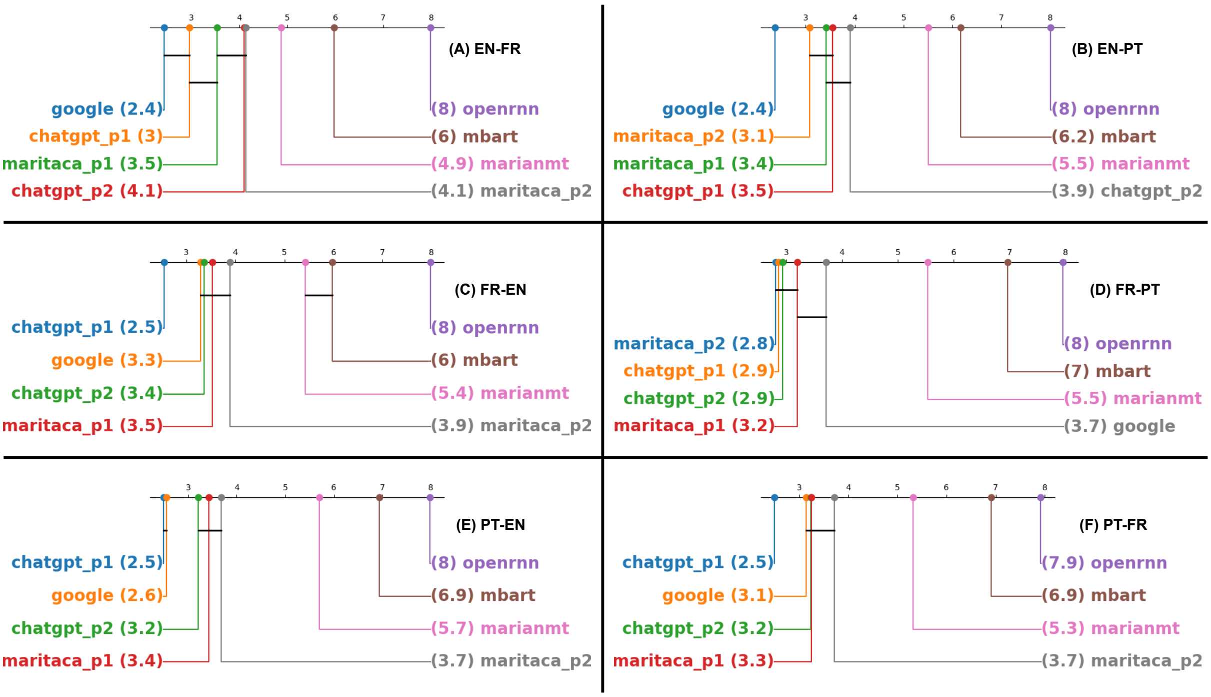
Table 9 – Friedman test results for translation metrics.

Language Pair	BLEU	METEOR	BERTScore
FR-EN	$p <$ 1e$-$228	$p <$ 1e$-$236	$p <$ 1e$-$243
FR-PT	$p \approx 0$	$p \approx 0$	$p \approx 0$
EN-FR	$p <$ 1e$-$237	$p <$ 1e$-$243	$p <$ 1e$-$235
EN-PT	$p <$ 1e$-$229	$p <$ 1e$-$239	$p <$ 1e$-$265
PT-FR	$p <$ 1e$-$235	$p <$ 1e$-$262	$p <$ 1e$-$294
PT-EN	$p <$ 1e$-$308	$p <$ 1e$-$306	$p \approx 0$

Source: Own authorship.

This statistical validation supports the use of post-hoc analyses with the Nemenyi test to identify which specific systems differ significantly from one another. The Nemenyi test performs pairwise comparisons among all models to detect statistically significant differences in performance based on their average rankings.

Figure 8 – BERTScore Critical Difference diagram for each translation.



Source: Own authorship.

Figure 8 presents the Critical Difference (CD) diagrams for all language pairs based on BERTScore, where lower ranks indicate better performance. The post-hoc analysis confirmed a consistent hierarchy among the evaluated systems, with LLMs and Google Translate forming the top-performing group across most translation directions, followed by MarianMT and mBART, and OpenNMT (RNN) significantly below all others.

There is a difference in the EN-PT pair. For BERTScore, Maritaca P2 performed better, while in the Nemenyi ranking, Google Translate came out on top. This can occur because BERTScore measures which model preserves semantic meaning better on average. The Nemenyi test also computes an average ranking, but its ranking reflects which model is more consistent in outperforming the others on a per-document basis. In other words, Maritaca P2 had some documents among the dataset where it performed exceptionally well, while Google maintained a slightly more consistent performance across all documents.

ChatGPT with prompt 1 exhibited stable and semantically faithful translations across directions, whereas Maritaca with prompt 2 excelled particularly in translations into Portuguese, benefiting from its linguistic familiarity. Google Translate performed competitively, especially in English-based pairs, supported by vast parallel corpora and continuous optimization. MarianMT and mBART achieved intermediate results, reflecting their smaller training scope and less adaptive architectures, though MarianMT generally outperformed mBART. OpenNMT with RNN, limited by sequential processing and reduced contextual capacity, consistently ranked last.

Overall, the Nemenyi test confirms that the differences observed in the scores are not due to random variation but reflect real performance gaps between systems. These findings reinforce the superiority of LLMs for creative language translation tasks, and also validate Google Translate as a strong baseline in this domain.

5.3.2 Evaluation Using Topic Modeling

The thematic analysis conducted with BERTopic is presented here. By comparing topics extracted from original poems with those present in reference translations and machine outputs, it becomes possible to assess the degree of thematic consistency preserved across translations in the baseline setup.

First, a search was conducted with different values of min_topic_size using BERTopic for each language pair in the original poems, in order to determine the configuration that achieved the best coherence score (c_v). The parameter min_topic_size defines the minimum number of documents required to form a topic, thus controlling the granularity of the clustering. The c_v coherence measure, in turn, evaluates the degree of semantic similarity between the most relevant words in a topic, serving as an indicator of topic interpretability. Values of min_topic_size ranging from 5 to 10 were tested. For each language pair, the configuration yielding the highest c_v score was selected. The results are presented in Table 10, which reports the chosen min_topic_size, the resulting number of clusters, and the corresponding c_v score.

Once the number of topics was defined from the original poems in each language pair, the poems were grouped accordingly. Based on this division of the originals, the translated poems from each model, along with the human reference translations, were

Table 10 – `min_topic_size` configuration for each language pair.

Language Pair	min_topic_size	#clusters	c_v
FR-EN	7	3	0.3936
FR-PT	7	4	0.3589
EN-FR	8	4	0.3414
EN-PT	9	3	0.3228
PT-FR	7	3	0.3711
PT-EN	6	2	0.3712

Source: Own authorship.

assigned to the same topics (clusters) as their corresponding originals. Subsequently, a list of representative words per topic was generated using c-TF-IDF, extracting the top 10 words for each topic in every language pair. These word lists were then used to create visual comparisons, allowing us to examine whether the most relevant words from the originals also appeared in the translations, thereby indicating whether thematic consistency was preserved.

In the visualizations, words displayed in the same colors represent exact matches across versions, while words in white indicate the absence of correspondence. The order of the words is also meaningful: c-TF-IDF ranks them in descending order of relevance, and this order is preserved in the graphs, with the most relevant terms appearing on the left. With these visualizations, it is possible to analyze each topic within each language pair. To illustrate the topic modeling, the French-English language pair was selected, as it achieved the highest *c_v* score, shown in Table 10, with 3 topics. Figure 9 illustrates topic 1; the other two topics for the FR-EN pair can be found in Appendix A, as well as the topics generated from the other language pairs.

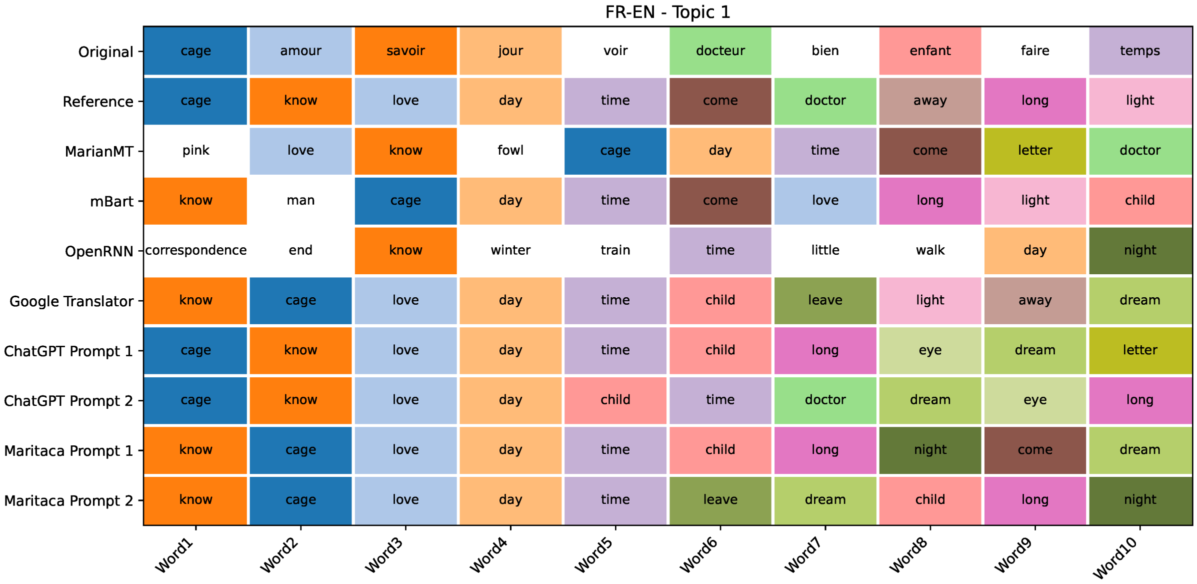
Tables 11 and 12 present the results of topic modeling for the translations produced by each system, comparing (i) the original poems, and (ii) the human reference translations. Each topic contains ten keywords, and each value indicates the number of words in which the system's translation shares the same dominant topic label as the corresponding comparison text The last column shows averaged across the three thematic clusters.

Table 11 – Topic overlap between system translations and the original poems.

Model	Topic 1	Topic 2	Topic 3	Mean
Reference	6	6	9	7.00
MarianMT	6	6	9	7.00
mBART	6	7	9	7.33
OpenNMT (RNN)	3	2	6	3.67
Google Translate	6	7	5	6.00
ChatGPT (Prompt 1)	6	6	9	7.00
ChatGPT (Prompt 2)	7	8	10	8.33
Maritaca (Prompt 1)	6	8	10	8.00
Maritaca (Prompt 2)	6	5	8	6.33

Source: Own authorship.

Figure 9 – Keywords for Topic 1 in the original poem and its translations for the French-English language pair.



FR-EN - Topic 1

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	cage	amour	savoir	jour	voir	docteur	bien	enfant	faire	temps
Reference	cage	know	love	day	time	come	doctor	away	long	light
MarianMT	pink	love	know	fowl	cage	day	time	come	letter	doctor
mBart	know	man	cage	day	time	come	love	long	light	child
OpenRNN	correspondence	end	know	winter	train	time	little	walk	day	night
Google Translator	know	cage	love	day	time	child	leave	light	away	dream
ChatGPT Prompt 1	cage	know	love	day	time	child	long	eye	dream	letter
ChatGPT Prompt 2	cage	know	love	day	child	time	doctor	dream	eye	long
Maritaca Prompt 1	know	cage	love	day	time	child	long	night	come	dream
Maritaca Prompt 2	know	cage	love	day	time	leave	dream	child	long	night

Source: Own authorship.

Table 12 – Topic overlap between system translations and human reference translations.

Model	Topic 1	Topic 2	Topic 3	Mean
MarianMT	7	7	8	7.33
mBART	8	6	8	7.33
OpenNMT (RNN)	3	3	6	4.00
Google Translate	7	8	5	6.67
ChatGPT (Prompt 1)	6	8	8	7.33
ChatGPT (Prompt 2)	7	6	9	7.33
Maritaca (Prompt 1)	7	8	9	8.00
Maritaca (Prompt 2)	6	6	7	6.33

Source: Own authorship.

In Table 11, the results reveal that the Reference – used only as a baseline to provide a notion of what constitutes a good number of overlapping keywords – MarianMT, and ChatGPT (Prompt 1) systems achieved identical average scores (7.0), and mBART slightly outperformed this group with a mean score of 7.33. Maritaca (Prompt 1) and ChatGPT (Prompt 2) reached the highest alignment scores (8.0 and 8.33, respectively), reflecting their ability to maintain thematic coherence while incorporating stylistic control. Conversely, OpenNMT (RNN) obtained the lowest mean score (3.67), confirming its limited capacity to preserve higher-level semantic structure and thematic continuity. Google Translate performed moderately (6.0), indicating reliable but less nuanced topic alignment compared to LLM-based models.

The fact that ChatGPT (Prompt 2) and mBART outperformed other systems in thematic alignment contrasts with the results obtained from the automatic evaluation metrics. While automatic metrics tend to reward lexical or embedding-level correspondence with the reference, topic modeling is more sensitive to whether the translated poem preserves the overall theme. mBART, as a multilingual denoising autoencoder trained on heterogeneous corpora, may therefore generalize better in maintaining global thematic structures, even when its lexical choices diverge. Similarly, ChatGPT (Prompt 2) often produced more creative outputs with lexical deviations, but still managed to preserve the central theme of the poem, which aligns with its higher topic-alignment score.

The reference-based comparison in Table 12 offers a complementary view, showing how closely each system reproduces the thematic structure found in human translations. Maritaca (Prompt 1) achieved the highest mean score (8.0), followed by MarianMT, mBART, ChatGPT (Prompt 1), and ChatGPT (Prompt 2), all tied at 7.33. Google Translate again occupied an intermediate position (6.67), while OpenNMT (RNN) remained well below the others (4.0).

The average scores obtained in the reference-based comparison are, in most cases, slightly higher than those in the original-based comparison. This indicates that the automatic translations share more keywords with the human reference translations than with the original poems, suggesting a convergence toward the lexical and thematic choices made by human translators rather than the exact content of the source text. This pattern can be explained by the tendency of both automatic and human translations to follow the linguistic and semantic conventions of the target language.

A closer examination of the word order within Topic 1 (Figure 9) reveals additional information about how each model captures the thematic hierarchy of the original poem. Since c-TF-IDF ranks terms by their relative importance within the topic, words positioned further to the left exert a stronger influence on its semantic core. Therefore, preserving their order indicates that the translation not only retains the same concepts but also maintains their relative thematic prominence. In the source text, the most salient words – “cage”, “amour”, and “savoir” – form a semantic nucleus that combines the notions of confinement, affection, and knowledge. The human reference maintains a comparable structure, although slightly altering the order by placing “know” before “love”.

Among the automatic systems, none reproduces the exact order of the original, but ChatGPT (both prompts) exhibits the closest match to the reference in both the selection and ordering of keywords: “cage”, “know”, and “love” remain the three most relevant terms. Maritaca (both prompts) and Google Translate display strong thematic consistency but slightly modify the initial order, placing “know” before “cage” and “love”, which alters the hierarchy found in both the reference and the original text, shifting the emphasis from the imagery of confinement toward cognition.

Marian preserves the relative order of “love” and “know” from the original, but “cage” appears in the middle. mBART changes the overall order, although “know” and “cage” remain among the three most salient words, while “love” moves toward the middle of the list. OpenNMT-RNN retains only “know”, with “cage” and “love” no longer appearing among the most relevant terms, indicating that its translation deviates more clearly from the original thematic core.

5.3.3 Human Evaluation

The human evaluation of the automatically translated poems focuses on aspects such as Poetic Structure, Poetic Stylistics, Fluency, Meaning, and Overall Impression. This perspective complements the quantitative findings with a literary-oriented assessment.

The evaluation criteria described in Section 4.4.3 were submitted to five specialists, who assigned scores to the automatic translations selected for each language pair. As previously noted, for each pair, two automatic translations were chosen: one produced by an LLM and another by a specialized MT system, both corresponding to the models that achieved the highest BERTScore. The evaluators received the original poem, the human reference translation, and the two automatic translations.

Table 13 presents the selected language pairs, poems, human reference versions, and the MT and LLM systems included in the evaluation. The poems were carefully selected to ensure that none of them were used in any training or fine-tuning processes. Appendix B.1 contains the poems used in this evaluation, along with their respective translations. Table 14 presents the average scores and standard deviations for each language pair and model. The bold line indicates the model achieving the highest score for each pair. The complete evaluator score tables are presented in Appendix B.2.

Table 13 – Poems, human references, and translation systems selected for human evaluation.

Language Pair	**Poem**	**Human Reference**	**Specialized MT**	**LLM**
FR-EN	*Clair de lune* (Paul Verlaine)	*Moonlight* (Norman R. Shapiro)	Google Translate	ChatGPT Prompt 1
FR-PT	*Demain, dès l’aube* (Victor Hugo)	*Amanhã, ao amanhecer* (JB Xavier)	Google Translate	Maritaca Prompt 2
EN-FR	*The Road Not Taken* (Robert Frost)	*Le chemin délaissé* (François Charton)	Google Translate	ChatGPT Prompt 1
EN-PT	*Do Not Go Gentle into That Good Night* (Dylan Thomas)	*Não entres mansamente nessa noite funda* (José Francisco Botelho)	Google Translate	Maritaca Prompt 2
PT-FR	*Traduzir-se* (Ferreira Gullar)	*Se traduire* (L. Gonçalves & D. Lamaison)	Google Translate	ChatGPT Prompt 1
PT-EN	*Canção do Exílio* (Gonçalves Dias)	*The Song of Exile* (Nelson Ascher)	Google Translate	ChatGPT Prompt 1

Source: Own authorship.

The human evaluation revealed consistent tendencies across the six language pairs. The LLMs (ChatGPT and Maritaca) consistently outperformed Google Translate in the criteria of Fluency, Meaning, and Overall Impression, producing translations that were

Table 14 – Average score (mean / standard deviation) from the human evaluation by language pair and model.

L. Pair	Model	Structure	Stylistics	Fluency	Meaning	Overall
FR-EN	Google Translate	1.6 / 0.49	1.4 / 0.49	2.8 / 0.45	2.8 / 0.45	2.0 / 0.00
	ChatGPT P1	**1.8 / 0.42**	**1.8 / 0.42**	**3.8 / 0.45**	**3.8 / 0.45**	**3.2 / 0.37**
FR-PT	Google Translate	1.6 / 0.49	1.4 / 0.49	2.8 / 0.45	3.0 / 0.00	2.2 / 0.45
	Maritaca P2	**1.8 / 0.42**	**1.8 / 0.42**	**3.8 / 0.45**	**3.8 / 0.45**	**3.2 / 0.37**
EN-FR	Google Translate	1.6 / 0.49	1.4 / 0.49	2.8 / 0.45	3.0 / 0.00	2.2 / 0.45
	ChatGPT P1	**1.8 / 0.42**	**1.8 / 0.42**	**4.0 / 0.00**	**3.8 / 0.45**	**3.4 / 0.37**
EN-PT	Google Translate	1.6 / 0.49	1.4 / 0.49	2.8 / 0.45	3.0 / 0.00	2.2 / 0.45
	Maritaca P2	**1.8 / 0.42**	**1.8 / 0.42**	**4.0 / 0.00**	**3.8 / 0.45**	**3.4 / 0.37**
PT-FR	Google Translate	1.6 / 0.49	1.4 / 0.49	2.8 / 0.45	3.0 / 0.00	2.2 / 0.45
	ChatGPT P1	**1.8 / 0.42**	**1.8 / 0.42**	**4.0 / 0.00**	**3.8 / 0.45**	**3.4 / 0.37**
PT-EN	Google Translate	1.6 / 0.49	1.6 / 0.49	2.8 / 0.45	3.0 / 0.00	2.2 / 0.45
	ChatGPT P1	**1.8 / 0.42**	**1.8 / 0.42**	**4.0 / 0.00**	**4.0 / 0.00**	**3.4 / 0.45**

Source: Own authorship.

generally more natural and stylistically coherent. However, a crucial finding was that all systems, without exception, performed poorly in the criteria of Poetic Structure and Poetic Style, receiving consistently low scores from the experts. This result highlights a fundamental challenge for both LLMs and MT systems in capturing formal aspects of poetry, such as rhythm, rhyme, meter, and stylistic nuance. Furthermore, the relatively low standard deviations across evaluators indicate a strong consensus regarding the comparative performance of the systems, reinforcing the reliability of the observed patterns. While LLMs demonstrated a clear advantage in fluency and semantic adequacy, both approaches remain limited in faithfully preserving the artistic and structural dimensions of poetry.

5.4 Results of Module 2 - Lyrics-Adapted

Module 2 explores whether models can benefit from fine-tuning on a lyrics corpus, given the linguistic and stylistic similarities between lyrics and poetry. The resulting models are tested on the same set of 300 poems as the baseline, and their performance is measured through automatic evaluation metrics.

The results of fine-tuning MT models on song lyrics are presented in Table 15 using BERTScore, while BLEU and METEOR results can be found in Appendix C. This analysis examines whether exposure to lyric-specific stylistic and rhythmic patterns enhances performance in poetry translation, with outcomes directly compared against the baseline models introduced in Module 1. The table reports fine-tuning results alongside baseline scores, with bold values highlighting the best result for each language pair. The number

Table 15 – BERTScore evaluation results for mBART, MarianMT, and OpenNMT (Lyrics Fine-Tuning vs. Pre-Trained).

Language Pair	**mBART**		**MarianMT**		**OpenNMT (RNN)**	
	Lyrics FT	Pre-Trained	Lyrics FT	Pre-Trained	Lyrics FT	Opus-Trained
FR-EN	0.8720	**0.9029**	0.8392	**0.9024**	0.6001	**0.7977**
FR-PT	**0.7884**	0.7051	0.5355	**0.7832**	0.4904	**0.6013**
EN-FR	0.7514	**0.7812**	0.7576	**0.7989**	0.4708	**0.5759**
EN-PT	0.7563	**0.7737**	0.5793	**0.7822**	0.4557	**0.5762**
PT-FR	**0.7269**	0.7045	0.6573	**0.7657**	0.4933	**0.6383**
PT-EN	**0.8689**	0.8628	0.8069	**0.8980**	0.6155	**0.8049**

Source: Own authorship.

of lyrics added for each language pair is detailed in Subsection 5.1.

Fine-tuning on lyrics for mBART improved notably for FR-PT (from 70.51% to 78.84%), PT-FR (70.45% to 72.69%), and PT-EN (86.28% to 86.89%). However, slight decreases were observed for FR-EN, EN-FR, and EN-PT. In contrast, both MarianMT and OpenNMT-RNN consistently exhibited lower performance after fine-tuning on lyrics, with decreases observed across all language pairs, indicating reduced lexical alignment and semantic fidelity compared to the original models.

The difference in performance between MarianMT and mBART, both Transformer-based, after fine-tuning on lyrics, can be attributed to the models' pre-training characteristics and their flexibility in adapting to new domains. MarianMT, being optimized for direct translation between language pairs, performs strongly in its pre-trained state (JUNCZYS-DOWMUNT et al., 2018). Fine-tuning it with lyrics may have disrupted the model's ability to preserve the structural and semantic nuances of poems, leading to lower scores compared to its pre-trained version.

In contrast, mBART benefits from its multilingual denoising pre-training and more flexible architecture (CHAKRABARTY; SAAKYAN; MURESAN, 2021). This allows it to incorporate stylistic and creative patterns from a specific lyrics dataset without losing general translation quality. The model is thus better able to transfer knowledge from lyrics to poetry, capturing elements that improve overall translation performance. This suggests that fine-tuning on a related creative text domain is more effective for models with higher adaptability, while models specialized in direct translation are more sensitive to domain shifts.

OpenNMT, with its RNN architecture, shows the least benefit from fine-tuning, as traditional RNNs have a limited capacity to model long-range dependencies and complex semantic structures. The observed decline in scores after fine-tuning suggests that exposure to domain-specific data can confuse the model, resulting in worse performance than when relying solely on its already constrained general knowledge.

It is important to note, however, that the lyrics corpus used for fine-tuning is not balanced across language pairs, as shown in Section 5.1. This imbalance may have influenced the learning dynamics of the models, particularly in underrepresented directions such as French-Portuguese (around 3,000 pairs), where the smaller amount of training data could limit the benefits of domain adaptation. This effect is noticeable in MarianMT, whose performance for FR-PT is the lowest among all fine-tuned directions (53.55%). However, this phenomenon is not observed in mBART or OpenNMT-RNN: for both, the lowest scores correspond to EN-PT (75.63% in mBart and 45.57% in OpenNMT-RNN) and EN-FR (75.14% in mBart and 47.08% in OpenNMT-RNN), which actually have a considerably larger number of examples (around 8,000 pairs).

The pairs PT-EN, FR-EN, and PT-FR, each with approximately 9,000 pairs, are the most represented in the lyrics dataset. Translations into English (87.20% FR-EN and 86.89% FR-PT in mBart) achieve the highest scores, while PT-FR (72.69% in mBart), despite having a similar number of fine-tuning examples, does not reach comparable performance. This suggests that the pre-training of the model plays a significant role, particularly for lower-resource language pairs, as fine-tuning alone is not sufficient to achieve the best results.

5.5 Results of Module 3 - Poems and Lyrics-Adapted

Module 3 focuses on evaluating three experimental configurations designed to assess the impact of domain adaptation on poetry translation: (1) The first configuration fine-tunes the models on 80% of the poem dataset, reserving the remaining 20% for testing; (2) the second configuration fine-tunes the models on a hybrid corpus that combines song lyrics with the same 80% poem split, aiming to determine whether the inclusion of stylistically related data enhances translation performance; (3) finally, the third configuration uses the pre-trained models without fine-tuning, which are re-evaluated on the same 20% poem test split to provide a direct comparison baseline. The results of these specialized MT systems are presented in Table 16, where the best-performing values for each language pair and metric are highlighted.

For mBART, the pre-trained model reaches the highest score for FR-EN (90.76%), whereas the hybrid Poems+Lyrics FT excels in FR-PT (79.84%) and PT-FR (77.23%). Poems FT achieves the top score for EN-FR (79.16%), EN-PT (78.25%), and PT-EN (87.53%). These results indicate that mBART's performance benefits from fine-tuning, whether on poems alone or in combination with lyrics, it generally enhances translation quality for mBart.

The fine-tuning on MarianMT, either Poems or Poems+Lyrics shows no significant improvements, except for a minor increase in EN-PT BERTScore (77.23% pre-trained to 77.73% poems FT). OpenNMT also shows no meaningful gains: Poems+Lyrics Fine-Tuning does not improve, and Poems Fine-Tuning had only minimal increases for EN-PT

Table 16 – BERTScore evaluation for mBART, MarianMT, and OpenNMT (RNN) translations with different fine-tuning strategies. Bold values indicate the best result for each model and language pair.

Model	Language Pair	Poems FT	Poems+Lyrics FT	Pre-Trained
mBART	FR-EN	0.8837	0.8385	**0.9076**
	FR-PT	0.7818	**0.7984**	0.7108
	EN-FR	**0.7916**	0.7867	0.7745
	EN-PT	**0.7825**	0.7791	0.7657
	PT-FR	0.7437	**0.7723**	0.7031
	PT-EN	**0.8753**	0.7905	0.8667
MarianMT	FR-EN	0.8665	0.7647	**0.9041**
	FR-PT	0.7402	0.7209	**0.7718**
	EN-FR	0.7775	0.7570	**0.8012**
	EN-PT	**0.7773**	0.7580	0.7723
	PT-FR	0.7374	0.6115	**0.7673**
	PT-EN	0.8535	0.6469	**0.8993**
OpenNMT (RNN)	FR-EN	0.7808	0.5369	**0.8034**
	FR-PT	0.6005	0.5838	**0.6008**
	EN-FR	**0.5863**	0.5291	0.5752
	EN-PT	0.5682	0.5605	**0.5693**
	PT-FR	0.6074	0.5482	**0.6314**
	PT-EN	0.7982	0.5468	**0.8030**

Source: Own authorship.

(57.52% pre-trained to 58.63%).

For comparison, the best result of each MT model – whether from poems-only fine-tuning, poems+lyrics fine-tuning, or the pre-trained version – was selected and evaluated against Google Translate and the LLMs using the same 20% subset of the poem dataset. Table 17 presents the highest BERTScore values. Among the LLMs, the best-performing configuration corresponds to ChatGPT Prompt 1 for FR-EN, EN-FR, and PT-FR, and to Maritaca Prompt 2 for FR-PT, EN-PT, and PT-EN. The complete BERTScore results for each LLM prompt, as well as the full BLEU and METEOR scores, are provided in Appendix D.

Table 17 – Best BERTScore results for poetry translations across specialized MT models and LLMs using the same test set.

Language Pair	mBART	MarianMT	Google Translate	OpenNMT	LLM
FR-EN	0.9076	0.9041	0.9253	0.8034	**0.9293**
FR-PT	0.7984	0.7718	0.8014	0.6008	**0.8340**
EN-FR	0.7916	0.8012	**0.8200**	0.5863	0.8105
EN-PT	0.7825	0.7773	0.7969	0.5693	**0.8183**
PT-FR	0.7723	0.7673	**0.8002**	0.6314	0.7997
PT-EN	0.8753	0.8993	**0.9256**	0.8030	0.9236

Source: Own authorship.

The results in Table 17 reveal that across all six language pairs, Google Translate achieves the highest scores among the specialized MT systems, ranging from 79.7% (EN-PT) to 92.6% (PT-EN). On average, it surpasses both mBART and MarianMT, confirming its overall robustness for poetic translation tasks. For instance, in the FR-EN direction, Google Translate achieves 92.5%, outperforming mBART (90.8%) and MarianMT (90.4%). This finding aligns with the results reported in Module 1, where Google Translate also outperformed the other MT models in the baseline evaluation using the full set of 300 poems. The same performance pattern is thus maintained, whether considering the 100% test set or the 20% test subset.

The LLMs achieve the highest scores in three of the six directions: FR-EN (92.9%), FR-PT (83.4%), and EN-PT (81.8%). Even in directions where Google Translate remains superior, the LLMs achieve very close results, with differences below 1%. While Google Translate maintains strong consistency across all directions, LLMs exhibit competitive or superior performance in most cases, suggesting a high degree of adaptability to poetic text.

As this analysis considers the best results obtained through fine-tuning, mBART tends to outperform MarianMT for most language pairs (FR-EN, FR-PT, EN-PT, PT-FR) when fine-tuned, whether on poems alone or on a combination of poems and lyrics. OpenNMT (RNN) consistently presents the lowest scores, with values between 56.9% (EN-PT) and 80.34% (FR-EN).

5.6 Final Considerations

The quantitative results, statistically validated with the Friedman test, revealed a clear performance hierarchy. LLMs such as ChatGPT-3.5 and Maritaca AI, along with Google Translate, consistently outperformed specialized MT models such as mBART and MarianMT. The Nemenyi post-hoc analysis confirmed these differences, highlighting the ability of LLMs and Google Translate to generate translations with higher semantic similarity. By contrast, the OpenNMT (RNN) model showed the weakest performance across all metrics, underscoring the limitations of recurrent architectures when dealing with the complexity of poetic language. Additionally, it was shown that prompt engineering significantly impacts translation quality in LLMs: more generic prompts generally had higher semantic fidelity for ChatGPT, whereas Maritaca AI proved more effective at integrating poetic instructions in a balanced manner.

Another consideration concerns the potential prior exposure of LLMs and Google Translate to some of the evaluated poems or their human translations during training. Since the data were collected from online sources and these systems are also trained on large-scale web data, there is a possibility that they may have encountered the original poems or their translations during training. However, preventing such exposure would require the use of non–public-domain poems that have never been translated, followed by large-scale manual translation.

The qualitative analysis deepened these findings. Human evaluation by specialists, based on a structured questionnaire, corroborated that the LLMs achieved higher scores in Fluency and Meaning Preservation. However, all models, without exception, showed significant difficulties in Poetic Structure and Poetic Style, consistently receiving low scores. This result indicates that while current technology can transmit semantic content, recreating form, rhythm, rhyme, and figurative language remains a challenge. Topic modeling analysis with BERTopic complemented this view, showing that the top-performing systems were more effective in maintaining thematic consistency. An interesting finding was the tendency of MT to align more closely with the themes of the human reference translation than with those of the original poem, suggesting that the models may be reproducing patterns from existing translations rather than generating a direct interpretation of the source text.

The fine-tuning experiments revealed different responses depending on model architecture. mBART proved capable of benefiting from domain-specific training data, particularly when trained exclusively on poems. With a larger and more diverse poetic corpus, fine-tuning could further enhance performance and lead to stronger results. In contrast, MarianMT and OpenNMT (RNN) did not show consistent gains, with their performance often decreasing after fine-tuning. This suggests that mBART's more flexible architecture and pre-training enable better adaptation to creative domains, while more specialized or simpler models may be more sensitive to domain shifts.

Chapter 6

Conclusion

The main objective of this work was to investigate how different machine learning models models handle the translation of poetry, a task that requires not only the transfer of literal meaning but also the preservation of stylistic, rhythmic, and expressive elements. The comparison was among specialized MT models (MarianMT, mBART, OpenNMT RNN, and Google Translate) and large language models (ChatGPT-3.5 and Maritaca AI), across six language pairs (FR-EN, FR-PT, EN-FR, EN-PT, PT-FR, PT-EN). To address this challenge, we proposed a three-phase evaluation framework integrating automatic metrics, thematic analysis, and expert human assessment. Furthermore, we examined the impact of different fine-tuning strategies on specialized translation models, using corpora composed of poems, song lyrics, and their combination, in order to evaluate whether adapting models to poetic language leads to measurable improvements in translation quality across multiple language pairs.

In line with the motivation, the results of this study reinforce the view of poetic translation as a particularly demanding test bed for machine translation. Unlike more utilitarian translation tasks, poetry requires sensitivity not only to semantic adequacy but also to rhythm, stylistic choices, and expressive nuance. As observed in the analyses, these characteristics expose limitations of purely automatic metrics and help explain qualitative differences between specialized MT models and LLMs.

The evaluation of Module 1, which establishes the baseline using pre-trained translation models evaluated through automatic metrics, topic modeling, and human assessment, demonstrates a clear performance hierarchy among translation models. Amidst the specialized MT models, Google Translate achieved the best performance across all language pairs and metrics, reaching 92.29% in BERTScore for the PT-EN pair. OpenNMT (RNN) consistently performed worst, and between mBART and MarianMT, the latter obtained the best results.However, even though Google Translate achieved highly competitive results, LLMs surpassed its performance when considering BERTScore, a metric particularly relevant for evaluating poetry. ChatGPT Prompt 1, the more generalist prompt, outperformed ChatGPT Prompt 2, which included instructions to preserve poetic characteristics, indicating that this prompt led to greater deviations. On the other hand, Maritaca AI Prompt 2 achieved better results in language pairs with translation into Portuguese, showing that its training on a Brazilian Portuguese corpus improved performance when translating into PT. Overall, translations into English, whether by LLMs or MT models, achieved the highest scores, suggesting that high-resource languages benefit training and lead to better results.

Topic modeling revealed a similar pattern with LLMs maintaining topic coherence better than specialized MT models. However, when comparing the persistence of keywords in automatic translations relative to the original and to the human reference translations, the latter showed better topic preservation. This suggests that MT models may rely more on existing references than on directly preserving the content of the original text. Human evaluation also supported the findings from automatic metrics, showing that LLMs better preserve Meaning and Fluency, although all models struggled to maintain Poetic Structure and Stylistic.

In particular, the inclusion of human evaluation proved essential for complementing the insights obtained from automatic and thematic analyses. While quantitative metrics primarily measure lexical and semantic similarity, human evaluation captures the nuanced aspects of poetic translation that automated measures cannot fully reflect. On one hand, it corroborated the results of automatic metrics by also identifying the superiority of LLMs in fluency and semantic preservation. On the other hand, it revealed the limitations of purely quantitative analyses by exposing gaps that thematic modeling and automatic metrics fail to detect. Human judgment was the only approach capable of critically assessing dimensions central to poetry, including the maintenance of poetic structure (meter, rhyme, stanza form) and style (figurative language, ambiguity, and tone). This qualitative analysis emphasized the importance of human judgment in evaluating creative texts, where aesthetic and interpretive dimensions extend beyond measurable linguistic correspondence.

Module 2 introduced fine-tuning using a corpus composed exclusively of song lyrics. For mBART, lyric-based fine-tuning improved scores in some language pairs, but for MarianMT and OpenNMT (RNN), no improvement was observed. Module 3 compared models fine-tuned exclusively on poems with models trained on a combination of poems and lyrics. The results showed that mBART benefited from both fine-tuning approaches, whereas MarianMT and OpenNMT performed better in their pretrained versions. Still, when compared to Google Translate and LLMs, fine-tuned MT models remained behind in overall scores.

The proposed evaluation framework proved effective in capturing different dimensions of translation quality, analysing semantic, thematic, and stylistic criteria. While Google Translate and LLMs currently offer strong performance in translation, they still struggle to preserve poetic characteristics, highlighting the ongoing challenges in automatic poetic translation.

As for the hypotheses, the findings support H_1, as large language models consistently outperformed traditional MT models across automatic metrics, thematic coherence, and human evaluations. Although Google Translate was the strongest among the specializaed MT models and achieved competitive results on automatic metrics such as BLEU and METEOR, the analysis using BERTScore, wich is particularly relevant for poetry due to its ability to capture contextual semantic similarity through embeddings, showed that large language models performed better across most language pairs. In contrast, H_2 was not supported, since fine-tuning with poems and song lyrics led to improvements only in a single model and in a limited number of language pairs, while most evaluated models showed no gains in performance.

6.1 Main Contributions

This work makes the following contributions:

- ❏ Development of a comprehensive evaluation framework for poetry translation that combines automatic metrics, topic-based analysis, and human expert assessment. This framework captures semantic, thematic, and stylistic dimensions.
- ❏ Application of topic modeling as an evaluator for thematic preservation in translated poetry.
- ❏ Development of a structured human evaluation questionnaire for literary translation.
- ❏ Empirical analysis of translation model performance.
- ❏ Investigation of the influence of language resources, showing that translations into English generally achieved higher scores, highlighting the benefit of high-resource languages in model training.
- ❏ Study of fine-tuning strategies using poems and song lyrics.
- ❏ Contribution to multilingual evaluation, focusing on Latin-based languages that are underrepresented in existing research.
- ❏ Creation of a multilingual poetry dataset comprising the six language pairs used in this study, including aligned original poems and human reference translations.
- ❏ Public release of the source code in the project repository to ensure transparency and reproducibility.

6.2 Future Work

Future research should aim to enhance MT models to more effectively capture the complexity of poetic form, including rhythm, meter, rhyme, and figurative language. To address this, researchers should develop more sophisticated architectures and training strategies that integrate explicit modeling of poetic constraints. Moreover, curating larger and more targeted corpora specifically designed for poetry, including diverse genres, authors, and their human translations, could provide models with richer examples of stylistic and rhythmic patterns, enabling better generalization across different poetic forms and language pairs.

Additionally, it would be beneficial to expand the number of language pairs analyzed, particularly including low-resource languages, since these languages present greater challenges in terms of available training data and translation quality, allowing for a more comprehensive assessment of model generalizability across diverse linguistic contexts. Further studies should also investigate a wider range of MT models and LLMs, evaluating their ability to handle underrepresented languages and complex poetic structures. Furthermore, expanding human evaluation by increasing both the number of poems analyzed and the number of expert evaluators would provide deeper insight into how models perform across different poetic styles, authors, and cultural contexts.

6.3 Contributions in Bibliographic Production

- The article “Comparative Analysis of Text Classification Algorithms” (BORGES; FARIA; GABRIEL, 2025), which presents a comparative study between Transformer-based models (BERT/BERTimbau) and traditional machine learning algorithms (Decision Tree, XGBoost, SVM, and MLP) for classification, was accepted at the XXII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC) 2025.

- The article “Comparison of Clustering Techniques in Portuguese Text Documents” (BORGES, 2025), which evaluates the performance of K-Means, Single Linkage, and Gaussian Mixture Model (GMM) algorithms for clustering Portuguese news texts using BERTimbau embeddings, was published in the iSys - Brazilian Journal of Information Systems.

- The article “Evaluating Translation Models and LLMs in Poetic Texts” (BORGES; GABRIEL; FARIA, 2025), which investigates the performance of machine translation models and LLMs in poetry translation across six language pairs, was accepted to the 18th International Conference on Agents and Artificial Intelligence (ICAART) 2026.

- This work, Between Rhymes and Algorithms: An Investigation into Poetic Machine Translation, was presented at the XIX Workshop of Theses and Dissertations in Computer Science in 2025 at the Federal University of Uberlândia (UFU). It received the Best Poster Award at the event.

Bibliography

BABALOLA, O.; OJOKOH, B.; BOYINBODE, O. Comprehensive evaluation of lda, nmf, and bertopic's performance on news headline topic modeling. **Journal of Computing and Theoretical Applications**, v. 2, n. 2, p. 268–289, Nov 2024.

BANERJEE, S.; LAVIE, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, Association for Computational Linguistics, Ann Arbor, Michigan, p. 65–72, jun 2005. Disponível em: <https://aclanthology.org/W05-0909>. Acesso em: 24 nov. 2024.

BICHRI, H.; CHERGUI, A.; HAIN, M. Investigating the impact of train / test split ratio on the performance of pre-trained models with custom datasets. **International Journal of Advanced Computer Science and Applications**, v. 15, n. 2, 2024.

BLEI, D. M. Probabilistic topic models. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 55, n. 4, p. 77–84, abr. 2012. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/2133806.2133826>.

BOASE-BEIER, J.; FISHER, L.; FURUKAWA, H. (Ed.). **The Palgrave Handbook of Literary Translation**. London: Palgrave Macmillan, 2018. 551 p. ISBN 9783319757520.

BORGES, B. **Multilingual Corpus of Poems with Human Reference Translations for Literary Machine Translation Research**. 2025. Mendeley Data. Disponível em: <https://doi.org/10.17632/gpmg8hdshh.1>.

BORGES, B.; FARIA, E.; GABRIEL, P. Comparative analysis of text classification algorithms. In: **Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional**. Porto Alegre, RS, Brasil: SBC, 2025. p. 189–200. ISSN 2763-9061. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/38718>.

BORGES, B. R. Comparison of clustering techniques in text documents in portuguese. **iSys - Brazilian Journal of Information Systems**, v. 18, n. 1, p. 4:1 – 4:17, Mar. 2025. Disponível em: <https://journals-sol.sbc.org.br/index.php/isys/article/view/5029>.

BRITTO, P. H. **A Tradução Literária**. Rio de Janeiro: Civilização Brasileira, 2012. 157 p. (Filosofia, Literatura e Artes). ISBN 978-85-200-1143-0.

BROWN, P.; COCKE, J.; PIETRA, S. D.; PIETRA, V. D.; JELINEK, F.; MERCER, R.; ROOSSIN, P. A statistical approach to language translation. **Proceedings of the International Conference on Computational Linguistics, Colling Budapest 1988, Volume 1**, 1988. Disponível em: <https://aclanthology.org/C88-1016>. Acesso em: 28 out. 2024.

CALLISON-BURCH, C.; OSBORNE, M.; KOEHN, P. Re-evaluating the role of Bleu in machine translation research. In: MCCARTHY, D.; WINTNER, S. (Ed.). **11th Conference of the European Chapter of the Association for Computational Linguistics**. Trento, Italy: Association for Computational Linguistics, 2006. p. 249–256. Disponível em: <https://aclanthology.org/E06-1032/>.

CANDIDO, A. **Literatura e Sociedade**. 9a edição, revista pelo autor. ed. Rio de Janeiro: Ouro sobre Azul, 2006. 200 p.

CASELI, H. d. M.; NUNES, M. d. G. V. **Processamento de linguagem natural: conceitos, técnicas e aplicações em português**. São Paulo: BPLN, 2023.

CHAKRABARTY, T.; SAAKYAN, A.; MURESAN, S. Don't go far off: An empirical study on neural poetry translation. In: MOENS, M.-F.; HUANG, X.; SPECIA, L.; YIH, S. W.-t. (Ed.). **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 7253–7265. Disponível em: <https://aclanthology.org/2021.emnlp-main.577>. Acesso em: 28 out. 2024.

CHATZIKOUMI, E. How to evaluate machine translation: A review of automated and human metrics. **Natural Language Engineering**, v. 26, n. 2, p. 137–161, 2020. Acesso em: 20 nov. 2024.

CONSTANTINE, P. Google translate gets voltaire: Literary translation and the age of artificial intelligence. **Contemporary French and Francophone Studies**, Routledge, v. 23, n. 4, p. 471–479, 2019. Disponível em: <https://doi.org/10.1080/17409292.2019.1694798>. Acesso em: 26 out. 2024.

GAO, R.; LIN, Y.; ZHAO, N. et al. Machine translation of chinese classical poetry: a comparison among chatgpt, google translate, and deepl translator. **Humanities and Social Sciences Communications**, v. 11, p. 835, 2024.

GARG, A.; AGARWAL, M. Machine translation: A literature review. **ArXiv**, abs/1901.01122, 2018. Disponível em: <https://doi.org/10.48550/arXiv.1901.01122>. Acesso em: 28 out. 2024.

Ghazvininejad, Marjan; Choi, Yejin; Knight, Kevin. Neural poetry translation. In: **Proceedings of NAACL-HLT**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 67–71.

GOLDSTEIN, N. **Versos, Sons, Ritmos**. 13ª edição. ed. São Paulo: Editora Ática, 2005. 80 p.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016. 800 p.

GROOTENDORST, M. Bertopic: neural topic modeling with a class-based tf-idf procedure. **arXiv preprint**, 2022. Disponível em: <https://arxiv.org/abs/2203.05794>. Acesso em: 24 nov. 2024.

HE, Z.; LIANG, T.; JIAO, W.; ZHANG, Z.; YANG, Y.; WANG, R.; TU, Z.; SHI, S.; WANG, X. Exploring human-like translation strategy with large language models. **Transactions of the Association for Computational Linguistics**, v. 12, p. 229–246, 03 2024. ISSN 2307-387X. Disponível em: <https://doi.org/10.1162/tacl_a_00642>. Acesso em: 02 dez. 2024.

HUMBLé, P. Machine translation and poetry. the case of english and portuguese. **Ilha do Desterro**, Universidade Federal de Santa Catarina, v. 72, n. 2, p. 41–56, May 2019. ISSN 2175-8026. Disponível em: <https://doi.org/10.5007/2175-8026.2019v72n2p41>. Acesso em: 28 out. 2024.

HUTCHINS, W. J.; SOMERS, H. L. **An Introduction to Machine Translation**. London: Academic Press Limited, 1992. 362 p. ISBN 0-12-362830-X.

JELODAR, H.; WANG, Y.; YUAN, C. et al. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. **Multimedia Tools and Applications**, v. 78, p. 15169–15211, 2019.

JONES, F. R. The translation of poetry. In: **The Oxford Handbook of Translation Studies**. Oxford University Press, 2011. p. 169–182. ISBN 9780199239306. Disponível em: <https://doi.org/10.1093/oxfordhb/9780199239306.013.0013>. Acesso em: 02 dez. 2024.

JUNCZYS-DOWMUNT, M.; GRUNDKIEWICZ, R.; DWOJAK, T.; HOANG, H.; HEAFIELD, K.; NECKERMANN, T.; SEIDE, F.; GERMANN, U.; AJI, A. F.; BOGOYCHEV, N.; MARTINS, A. F. T.; BIRCH, A. Marian: Fast neural machine translation in c++. **Proceedings of ACL 2018, System Demonstrations**, Association for Computational Linguistics, Melbourne, Australia, p. 116–121, jul 2018. Disponível em: <https://aclanthology.org/P18-4020>. Acesso em: 20 nov. 2024.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. 3rd. ed. Hoboken, NJ: Pearson, 2024. 599 p. Online manuscript released August 20, 2024. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

KARAIVANOV, S.; RAYCHEV, V.; VECHEV, M. Phrase-based statistical translation of programming languages. In: **Proceedings of the 2014 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming & Software**. New York, NY, USA: Association for Computing Machinery, 2014. (Onward! 2014), p. 173–184. ISBN 9781450332101. Disponível em: <https://doi.org/10.1145/2661136.2661148>. Acesso em: 28 out. 2024.

KHAN, M.; ULLAH, K.; ALHARBI, Y.; ALFERAIDI, A.; ALHARBI, T. S.; YADAV, K.; ALSHARABI, N.; AHMAD, A. Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archive. **Applied Sciences**, v. 13, n. 15, 2023. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/13/15/8566>.

KLEIN, G.; KIM, Y.; DENG, Y.; SENELLART, J.; RUSH, A. Opennmt: Open-source toolkit for neural machine translation. **Proceedings of ACL 2017, System Demonstrations**, Association for Computational Linguistics, Vancouver, Canada, p. 67–72, jul 2017. Disponível em: <https://aclanthology.org/P17-4012>. Acesso em: 18 dez. 2024.

KOEHN, P. **Neural Machine Translation**. Cambridge University Press, 2020. Online publication date: May 2020. ISBN 9781108608480. Disponível em: <https://doi.org/10.1017/9781108608480>.

LIDDICOAT, A. J. Intercultural mediation, intercultural communication and translation. **Perspectives: Studies in Translatology**, Taylor & Francis, v. 24, n. 3, p. 354–364, 2016. Acesso em: 12 set. 2025.

MADKOUR, M. Linguistic levels of translation: A generic exploration of translation difficulties in literary textual corpus. **International Journal of Applied Linguistics and English Literature**, v. 5, p. 99–118, 2016. Disponível em: <https://api.semanticscholar.org/CorpusID:55734423>. Acesso em: 20 nov. 2024.

MANNING, C. D.; SCHüTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, MA: MIT Press, 1999.

NAVARRO-COLORADO, B. On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry. **Frontiers in Digital Humanities**, v. 5, 2018. Disponível em: <https://doi.org/10.3389/fdigh.2018.00015>. Acesso em: 24 nov. 2024.

NORD, C.; ALMEIDA, H. do C.; ABREU, J. de; ZIPSER, M. E.; AIO, M. de A.; POLCHLOPEK, S. A. **Análise textual em tradução: bases teóricas, métodos e aplicação didática**. 1ª edição brasileira. ed. São Paulo, Brasil: Rafael Zamperetti Copetti Editor Ltda, 2016. 456 p. (Coleção Transtextos, Vol. 01, 1ª Série). ISBN 978-85-67569-26-0.

OZTURK, E. Preservation of cultural authenticity: Analysing cultural transfer in the translation process. **International Journal of Applied Linguistics and Translation**, v. 10, n. 4, p. 61–66, 2024. Disponível em: <https://doi.org/10.11648/j.ijalt.20241004.12>. Acesso em: 13 dez. 2024.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. USA: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. Disponível em: <https://doi.org/10.3115/1073083.1073135>. Acesso em: 23 nov. 2024.

PARKS, T. **Translating Style: A Literary Approach to Translation - A Translation Approach to Literature**. Manchester: St. Jerome Publishing / Routledge, 2007. 268 p. ISBN 9781905763047.

PIRES, R.; ABONIZIO, H.; ALMEIDA, T. S.; NOGUEIRA, R. Sabiá: Portuguese large language models. **Proceedings of the 12th Brazilian Conference on Intelligent Systems (BRACIS)**, Sociedade Brasileira de Computação, Belo Horizonte/MG, p. 226–240, 2023. ISSN 2643-6264.

Poe, E. A. O corvo. In: **Edgar Allan Poe: Medo Clássico. Vol. 1**. 1. ed. Rio de Janeiro: Darkside Books, 2017. ISBN 978-8594540249. Poema traduzido por Fernando Pessoa.

RESENDE, N.; HADLEY, J. The translator's canvas: Using llms to enhance poetry translation. In: KNOWLES, R.; ERIGUCHI, A.; GOEL, S. (Ed.). **Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)**. Chicago, USA: Association for Machine Translation in the Americas, 2024. p. 178–189. Disponível em: <https://aclanthology.org/2024.amta-research.16>. Acesso em: 28 out. 2024.

ROEDER, T.; BOTH, A.; HINNEBURG, A. Evaluating topic coherence measures. In: ACM. **Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM 2015)**. Shanghai, China, 2015. p. 399–408. Disponível em: <https://doi.org/10.1145/2684822.2685324>. Acesso em: 29 Oct 2025.

RUSSELL, S. J.; NORVIG, P. **Inteligência Artificial**. 3. ed. Rio de Janeiro: Elsevier, 2013. 1016 p. ISBN 978-85-352-3701-6.

SONG, W. L.; XU, H.; WONG, D. F.; ZHAN, R.; CHAO, L. S.; WANG, S. Towards zero-shot multilingual poetry translation. In: UTIYAMA, M.; WANG, R. (Ed.). **Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track**. Macau SAR, China: Asia-Pacific Association for Machine Translation, 2023. p. 324–335. Disponível em: <https://aclanthology.org/2023.mtsummit-research.27>. Acesso em: 28 out. 2024.

STAHLBERG, F. Neural machine translation: A review. **Journal of Artificial Intelligence Research**, v. 69, p. 343–418, out. 2020. Disponível em: <https://doi.org/10.1613/jair.1.12007>. Acesso em: 14 nov. 2024.

TANG, Y.; TRAN, C.; LI, X.; CHEN, P.-J.; GOYAL, N.; CHAUDHARY, V.; GU, J.; FAN, A. Multilingual translation with extensible multilingual pretraining and finetuning. **ArXiv**, abs/2008.00401, 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:220936592>. Acesso em: 20 nov. 2024.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. In: **Advances in Neural Information Processing Systems**. Cambridge, MA: MIT Press, 2017. p. 6000–6010. Acesso em: 15 nov. 2024.

VELLOSO, M. P. **História & Modernismo**. 1ª. ed. Belo Horizonte, Brasil: Autêntica, 2010. 128 p. ISBN 978-8575264799.

VENUTI, L. **The Translator's Invisibility: A History of Translation**. London and New York: Routledge, 1995. 344 p. ISBN 9780415394550.

VáZQUEZ, A. I. C.; MITKOV, R. Machine translation of literary texts: Genres, times and systems. In: GUTIéRREZ, R. L.; PAREJA, A.; MITKOV, R. (Ed.). **Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications**. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, 2023. p. 48–53. Disponível em: <https://aclanthology.org/2023.nlp4tia-1.7>. Acesso em: 28 out. 2024.

WANG, Q.; LI, B.; XIAO, T.; ZHU, J.; LI, C.; WONG, D. F.; CHAO, L. S. Learning deep transformer models for machine translation. **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, Association for Computational Linguistics, Florence, Italy, p. 1810–1822, jul 2019. Disponível em: <https://aclanthology.org/P19-1176>. Acesso em: 13 dez. 2024.

WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q. V.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K.; KLINGNER, J.; SHAH, A.; JOHNSON, M.; LIU, X.; KAISER, L.; GOUWS, S.; KATO, Y.; KUDO, T.; KAZAWA, H.; STEVENS, K.; KURIAN, G.; PATIL, N.; WANG, W.; YOUNG, C.; SMITH, J. R.; RIESA, J.; RUDNICK, A.; VINYALS, O.; CORRADO, G. S.; HUGHES, M.; DEAN, J. Google's neural machine translation system: Bridging the gap between human and machine translation. **ArXiv**, abs/1609.08144, 2016. Disponível em: <https://api.semanticscholar.org/CorpusID:3603249>. Acesso em: 15 nov. 2024.

ZHANG, T.; KISHORE, V.; WU, F.; WEINBERGER, K. Q.; ARTZI, Y. Bertscore: Evaluating text generation with bert. In: OPENREVIEW.NET. **8th International Conference on Learning Representations (ICLR 2020)**. Addis Ababa, Ethiopia, 2020. <https://openreview.net/forum?id=SkeHuCVFDr>. Disponível em: <https://openreview.net/forum?id=SkeHuCVFDr>. Acesso em: 25 nov. 2024.

Appendix

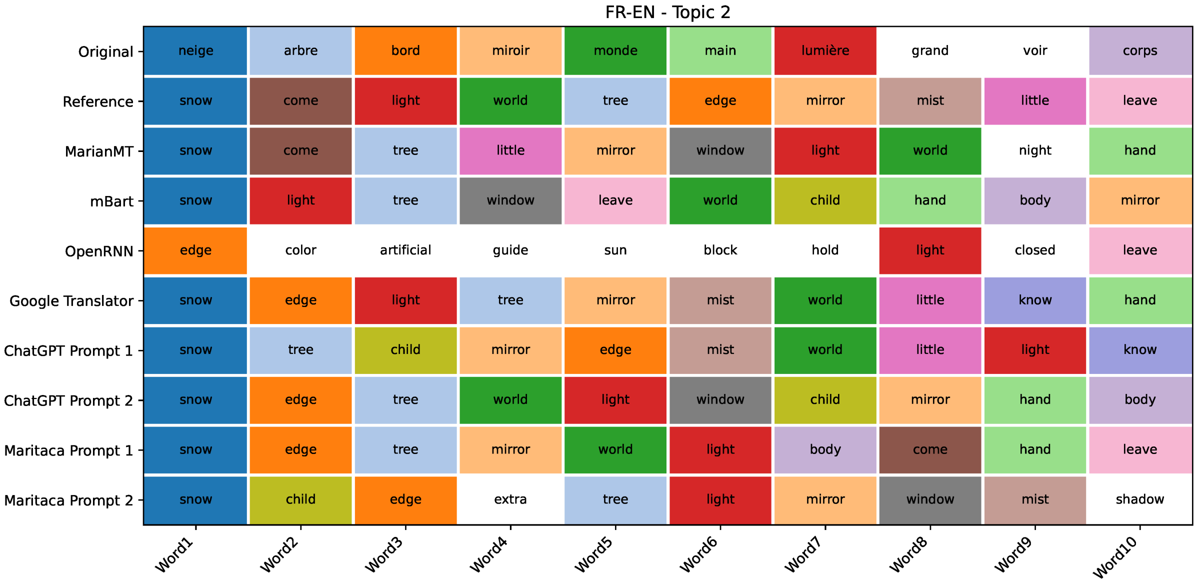
APPENDIX A

Topic Modeling Figures

This appendix presents the figures generated for the topic modeling analysis using BERTopic. In the visualizations, matching words between the original poem and its translations are represented by identical colors, while white indicates the absence of correspondence. The words are ordered from left to right according to their relevance within each topic. Figures are provided for all six language pairs and for every topic generated.

French-English

Figure 10 – Keywords for Topic 2 in the original poem and its translations for the French-English language pair.

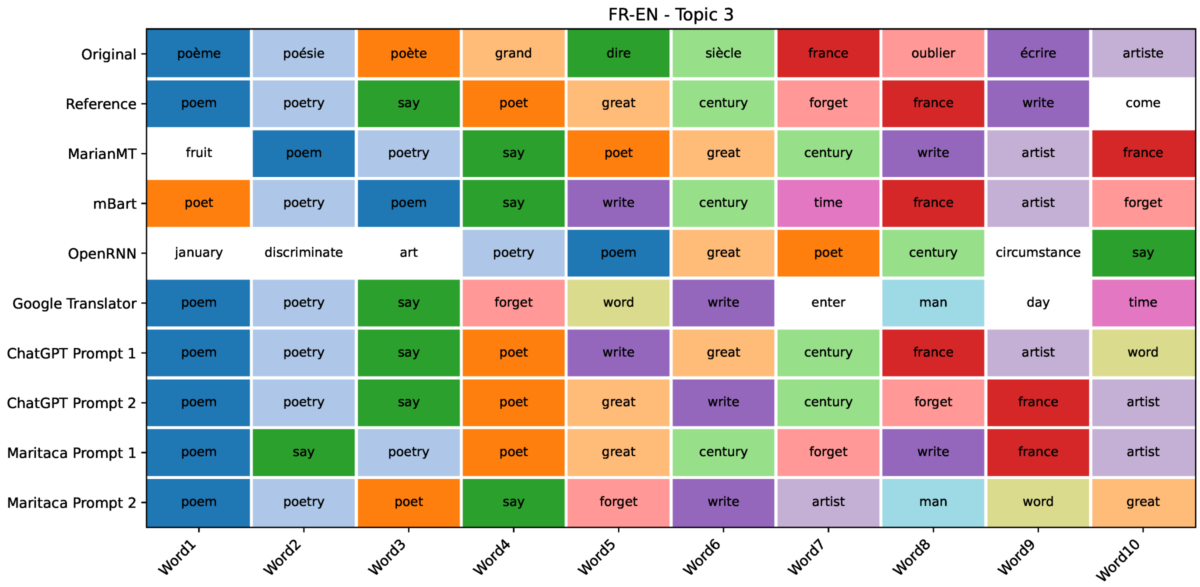


FR-EN - Topic 2

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	neige	arbre	bord	miroir	monde	main	lumière	grand	voir	corps
Reference	snow	come	light	world	tree	edge	mirror	mist	little	leave
MarianMT	snow	come	tree	little	mirror	window	light	world	night	hand
mBart	snow	light	tree	window	leave	world	child	hand	body	mirror
OpenRNN	edge	color	artificial	guide	sun	block	hold	light	closed	leave
Google Translator	snow	edge	light	tree	mirror	mist	world	little	know	hand
ChatGPT Prompt 1	snow	tree	child	mirror	edge	mist	world	little	light	know
ChatGPT Prompt 2	snow	edge	tree	world	light	window	child	mirror	hand	body
Maritaca Prompt 1	snow	edge	tree	mirror	world	light	body	come	hand	leave
Maritaca Prompt 2	snow	child	edge	extra	tree	light	mirror	window	mist	shadow

Source: Own authorship.

Figure 11 – Keywords for Topic 3 in the original poem and its translations for the French-English language pair.



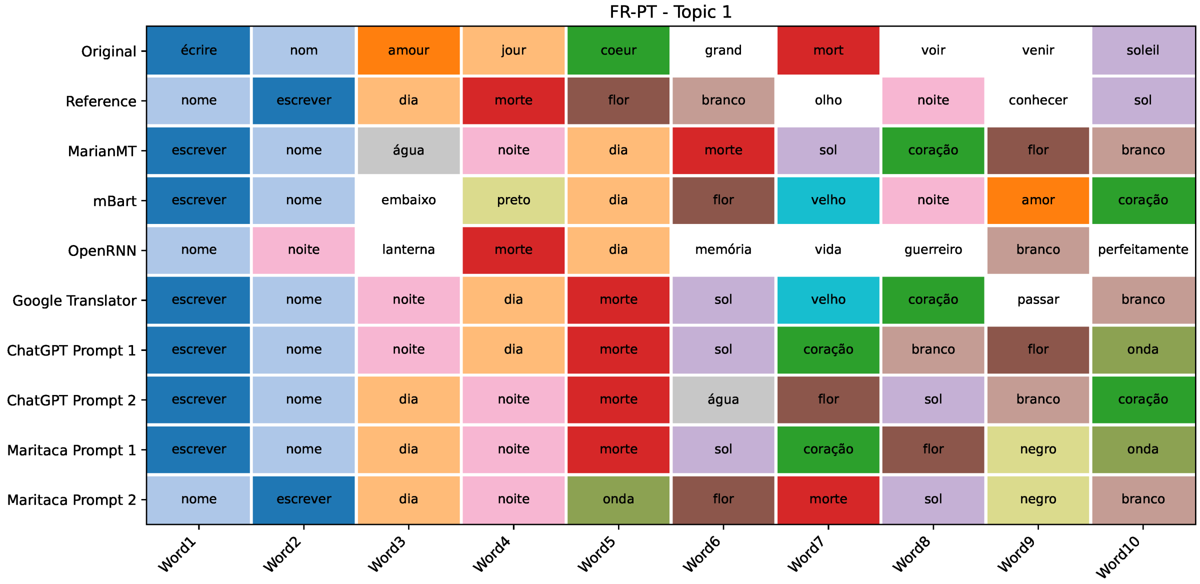
FR-EN - Topic 3

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	poème	poésie	poète	grand	dire	siècle	france	oublier	écrire	artiste
Reference	poem	poetry	say	poet	great	century	forget	france	write	come
MarianMT	fruit	poem	poetry	say	poet	great	century	write	artist	france
mBart	poet	poetry	poem	say	write	century	time	france	artist	forget
OpenRNN	january	discriminate	art	poetry	poem	great	poet	century	circumstance	say
Google Translator	poem	poetry	say	forget	word	write	enter	man	day	time
ChatGPT Prompt 1	poem	poetry	say	poet	write	great	century	france	artist	word
ChatGPT Prompt 2	poem	poetry	say	poet	great	write	century	forget	france	artist
Maritaca Prompt 1	poem	say	poetry	poet	great	century	forget	write	france	artist
Maritaca Prompt 2	poem	poetry	poet	say	forget	write	artist	man	word	great

Source: Own authorship.

French-Portuguese

Figure 12 – Keywords for Topic 1 in the original poem and its translations for the French-Portuguese language pair.

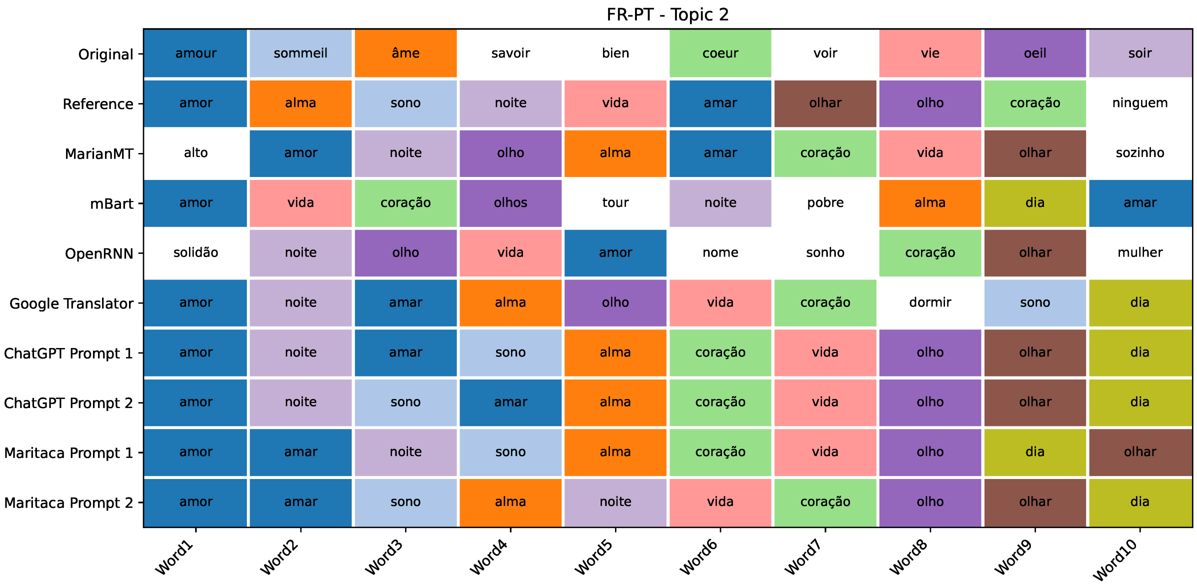


FR-PT - Topic 1

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	écrire	nom	amour	jour	coeur	grand	mort	voir	venir	soleil
Reference	nome	escrever	dia	morte	flor	branco	olho	noite	conhecer	sol
MarianMT	escrever	nome	água	noite	dia	morte	sol	coração	flor	branco
mBart	escrever	nome	embaixo	preto	dia	flor	velho	noite	amor	coração
OpenRNN	nome	noite	lanterna	morte	dia	memória	vida	guerreiro	branco	perfeitamente
Google Translator	escrever	nome	noite	dia	morte	sol	velho	coração	passar	branco
ChatGPT Prompt 1	escrever	nome	noite	dia	morte	sol	coração	branco	flor	onda
ChatGPT Prompt 2	escrever	nome	dia	noite	morte	água	flor	sol	branco	coração
Maritaca Prompt 1	escrever	nome	dia	noite	morte	sol	coração	flor	negro	onda
Maritaca Prompt 2	nome	escrever	dia	noite	onda	flor	morte	sol	negro	branco

Source: Own authorship.

Figure 13 – Keywords for Topic 2 in the original poem and its translations for the French-Portuguese language pair.

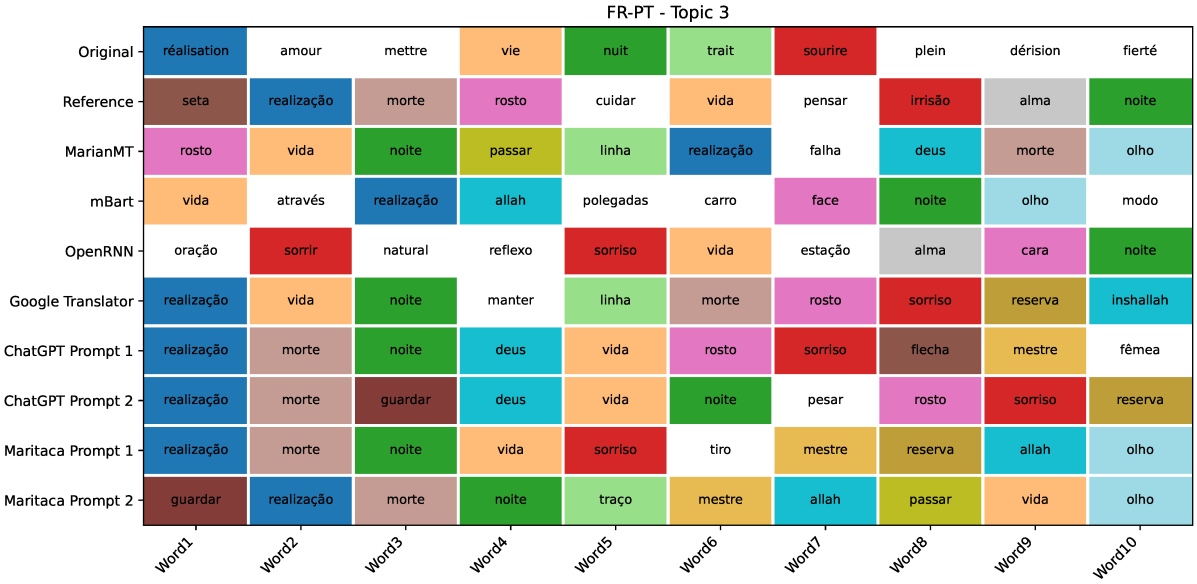


FR-PT - Topic 2

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	amour	sommeil	âme	savoir	bien	coeur	voir	vie	oeil	soir
Reference	amor	alma	sono	noite	vida	amar	olhar	olho	coração	ninguem
MarianMT	alto	amor	noite	olho	alma	amar	coração	vida	olhar	sozinho
mBart	amor	vida	coração	olhos	tour	noite	pobre	alma	dia	amar
OpenRNN	solidão	noite	olho	vida	amor	nome	sonho	coração	olhar	mulher
Google Translator	amor	noite	amar	alma	olho	vida	coração	dormir	sono	dia
ChatGPT Prompt 1	amor	noite	amar	sono	alma	coração	vida	olho	olhar	dia
ChatGPT Prompt 2	amor	noite	sono	amar	alma	coração	vida	olho	olhar	dia
Maritaca Prompt 1	amor	amar	noite	sono	alma	coração	vida	olho	dia	olhar
Maritaca Prompt 2	amor	amar	sono	alma	noite	vida	coração	olho	olhar	dia

Source: Own authorship.

Figure 14 – Keywords for Topic 3 in the original poem and its translations for the French-Portuguese language pair.

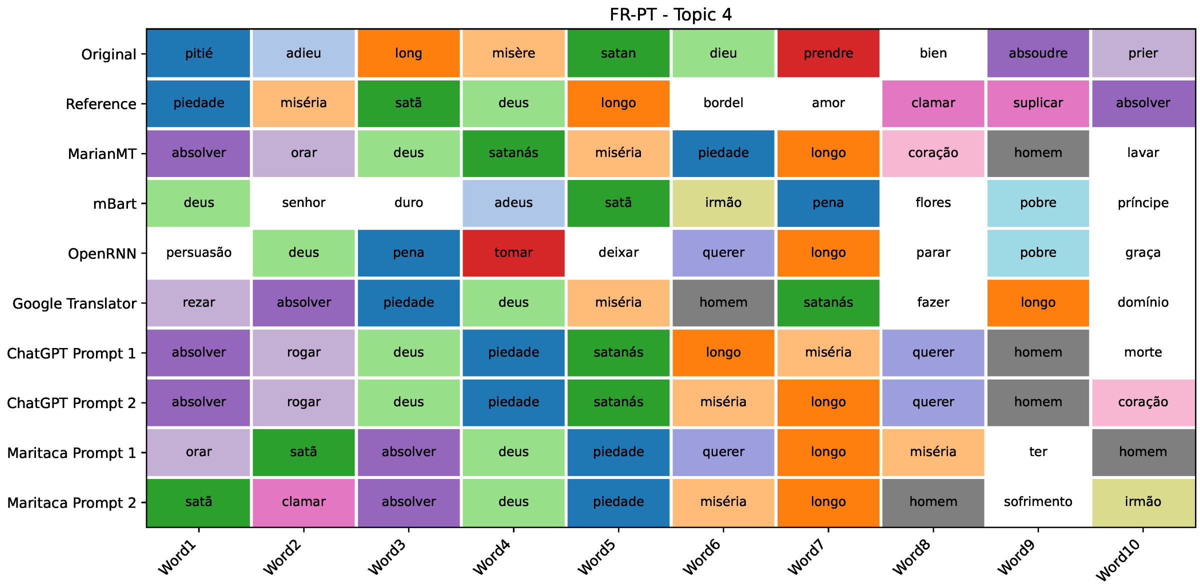


FR-PT - Topic 3

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	réalisation	amour	mettre	vie	nuit	trait	sourire	plein	dérision	fierté
Reference	seta	realização	morte	rosto	cuidar	vida	pensar	irrisão	alma	noite
MarianMT	rosto	vida	noite	passar	linha	realização	falha	deus	morte	olho
mBart	vida	através	realização	allah	polegadas	carro	face	noite	olho	modo
OpenRNN	oração	sorrir	natural	reflexo	sorriso	vida	estação	alma	cara	noite
Google Translator	realização	vida	noite	manter	linha	morte	rosto	sorriso	reserva	inshallah
ChatGPT Prompt 1	realização	morte	noite	deus	vida	rosto	sorriso	flecha	mestre	fêmea
ChatGPT Prompt 2	realização	morte	guardar	deus	vida	noite	pesar	rosto	sorriso	reserva
Maritaca Prompt 1	realização	morte	noite	vida	sorriso	tiro	mestre	reserva	allah	olho
Maritaca Prompt 2	guardar	realização	morte	noite	traço	mestre	allah	passar	vida	olho

Source: Own authorship.

Figure 15 – Keywords for Topic 4 in the original poem and its translations for the French-Portuguese language pair.



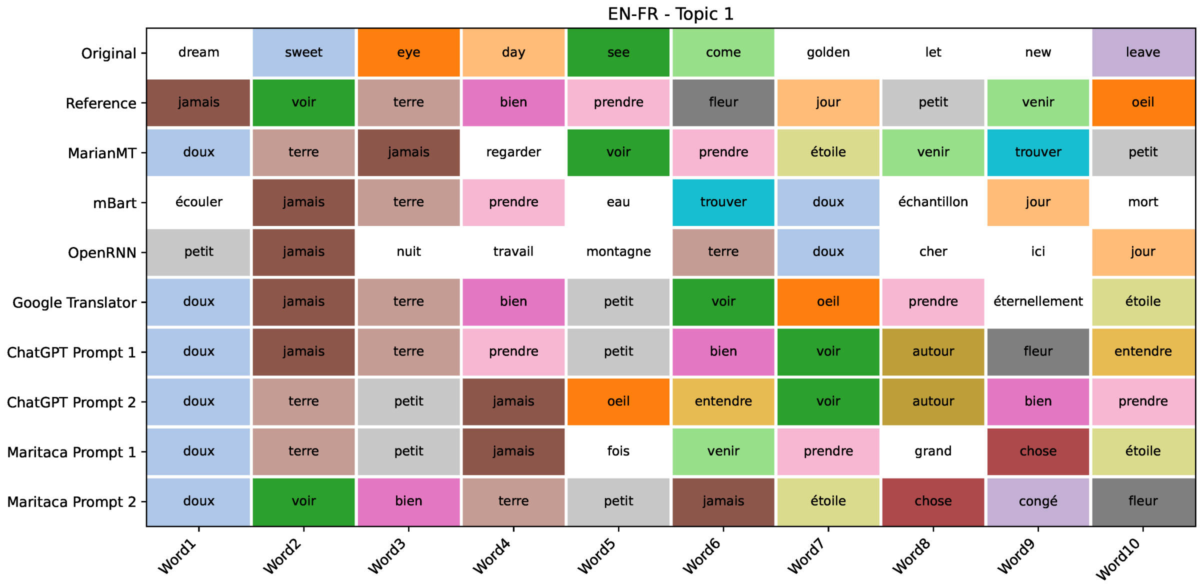
FR-PT - Topic 4

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	pitié	adieu	long	misère	satan	dieu	prendre	bien	absoudre	prier
Reference	piedade	miséria	satã	deus	longo	bordel	amor	clamar	suplicar	absolver
MarianMT	absolver	orar	deus	satanás	miséria	piedade	longo	coração	homem	lavar
mBart	deus	senhor	duro	adeus	satã	irmão	pena	flores	pobre	príncipe
OpenRNN	persuasão	deus	pena	tomar	deixar	querer	longo	parar	pobre	graça
Google Translator	rezar	absolver	piedade	deus	miséria	homem	satanás	fazer	longo	domínio
ChatGPT Prompt 1	absolver	rogar	deus	piedade	satanás	longo	miséria	querer	homem	morte
ChatGPT Prompt 2	absolver	rogar	deus	piedade	satanás	miséria	longo	querer	homem	coração
Maritaca Prompt 1	orar	satã	absolver	deus	piedade	querer	longo	miséria	ter	homem
Maritaca Prompt 2	satã	clamar	absolver	deus	piedade	miséria	longo	homem	sofrimento	irmão

Source: Own authorship.

English-French

Figure 16 – Keywords for Topic 1 in the original poem and its translations for the English-French language pair.

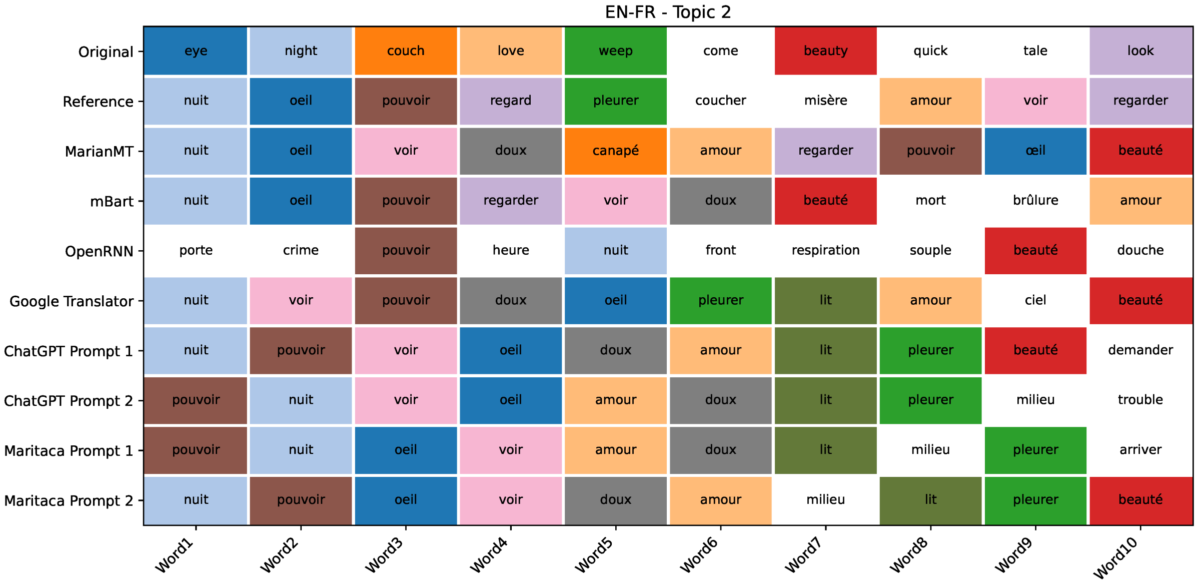


EN-FR - Topic 1

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	dream	sweet	eye	day	see	come	golden	let	new	leave
Reference	jamais	voir	terre	bien	prendre	fleur	jour	petit	venir	oeil
MarianMT	doux	terre	jamais	regarder	voir	prendre	étoile	venir	trouver	petit
mBart	écouler	jamais	terre	prendre	eau	trouver	doux	échantillon	jour	mort
OpenRNN	petit	jamais	nuit	travail	montagne	terre	doux	cher	ici	jour
Google Translator	doux	jamais	terre	bien	petit	voir	oeil	prendre	éternellement	étoile
ChatGPT Prompt 1	doux	jamais	terre	prendre	petit	bien	voir	autour	fleur	entendre
ChatGPT Prompt 2	doux	terre	petit	jamais	oeil	entendre	voir	autour	bien	prendre
Maritaca Prompt 1	doux	terre	petit	jamais	fois	venir	prendre	grand	chose	étoile
Maritaca Prompt 2	doux	voir	bien	terre	petit	jamais	étoile	chose	congé	fleur

Source: Own authorship.

Figure 17 – Keywords for Topic 2 in the original poem and its translations for the English-French language pair.

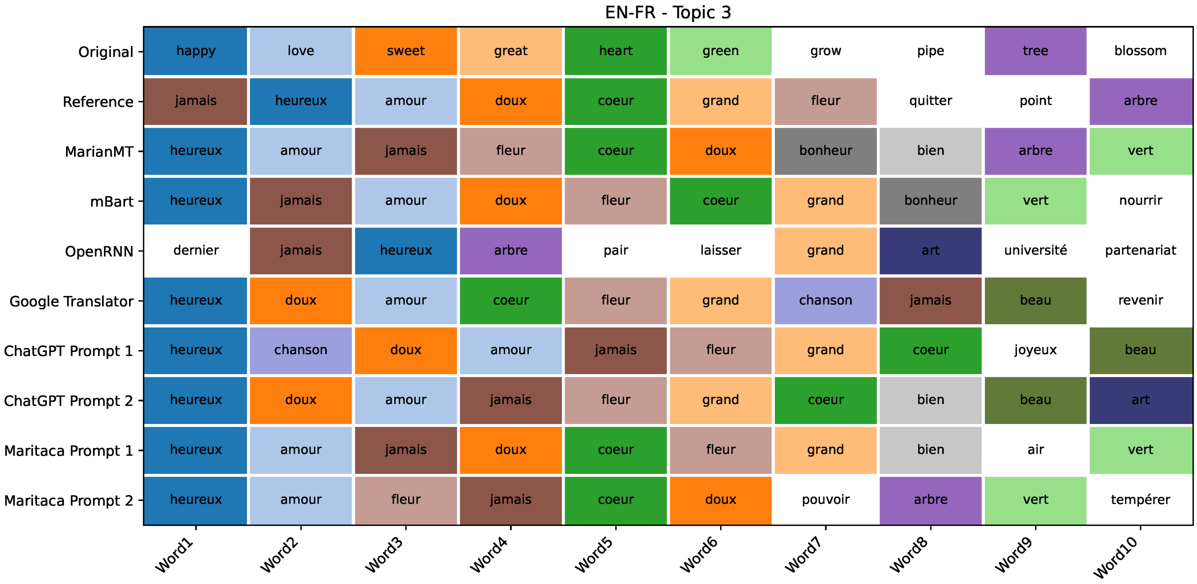


EN-FR - Topic 2

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	eye	night	couch	love	weep	come	beauty	quick	tale	look
Reference	nuit	oeil	pouvoir	regard	pleurer	coucher	misère	amour	voir	regarder
MarianMT	nuit	oeil	voir	doux	canapé	amour	regarder	pouvoir	œil	beauté
mBart	nuit	oeil	pouvoir	regarder	voir	doux	beauté	mort	brûlure	amour
OpenRNN	porte	crime	pouvoir	heure	nuit	front	respiration	souple	beauté	douche
Google Translator	nuit	voir	pouvoir	doux	oeil	pleurer	lit	amour	ciel	beauté
ChatGPT Prompt 1	nuit	pouvoir	voir	oeil	doux	amour	lit	pleurer	beauté	demander
ChatGPT Prompt 2	pouvoir	nuit	voir	oeil	amour	doux	lit	pleurer	milieu	trouble
Maritaca Prompt 1	pouvoir	nuit	oeil	voir	amour	doux	lit	milieu	pleurer	arriver
Maritaca Prompt 2	nuit	pouvoir	oeil	voir	doux	amour	milieu	lit	pleurer	beauté

Source: Own authorship.

Figure 18 – Keywords for Topic 3 in the original poem and its translations for the English-French language pair.

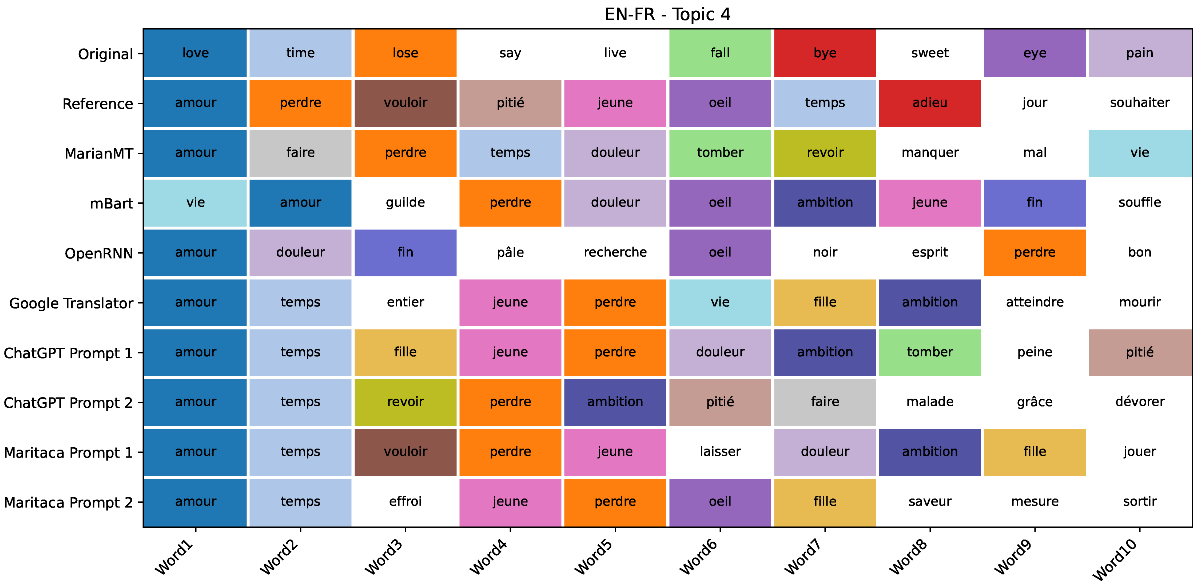


EN-FR - Topic 3

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	happy	love	sweet	great	heart	green	grow	pipe	tree	blossom
Reference	jamais	heureux	amour	doux	coeur	grand	fleur	quitter	point	arbre
MarianMT	heureux	amour	jamais	fleur	coeur	doux	bonheur	bien	arbre	vert
mBart	heureux	jamais	amour	doux	fleur	coeur	grand	bonheur	vert	nourrir
OpenRNN	dernier	jamais	heureux	arbre	pair	laisser	grand	art	université	partenariat
Google Translator	heureux	doux	amour	coeur	fleur	grand	chanson	jamais	beau	revenir
ChatGPT Prompt 1	heureux	chanson	doux	amour	jamais	fleur	grand	coeur	joyeux	beau
ChatGPT Prompt 2	heureux	doux	amour	jamais	fleur	grand	coeur	bien	beau	art
Maritaca Prompt 1	heureux	amour	jamais	doux	coeur	fleur	grand	bien	air	vert
Maritaca Prompt 2	heureux	amour	fleur	jamais	coeur	doux	pouvoir	arbre	vert	tempérer

Source: Own authorship.

Figure 19 – Keywords for Topic 4 in the original poem and its translations for the English-French language pair.

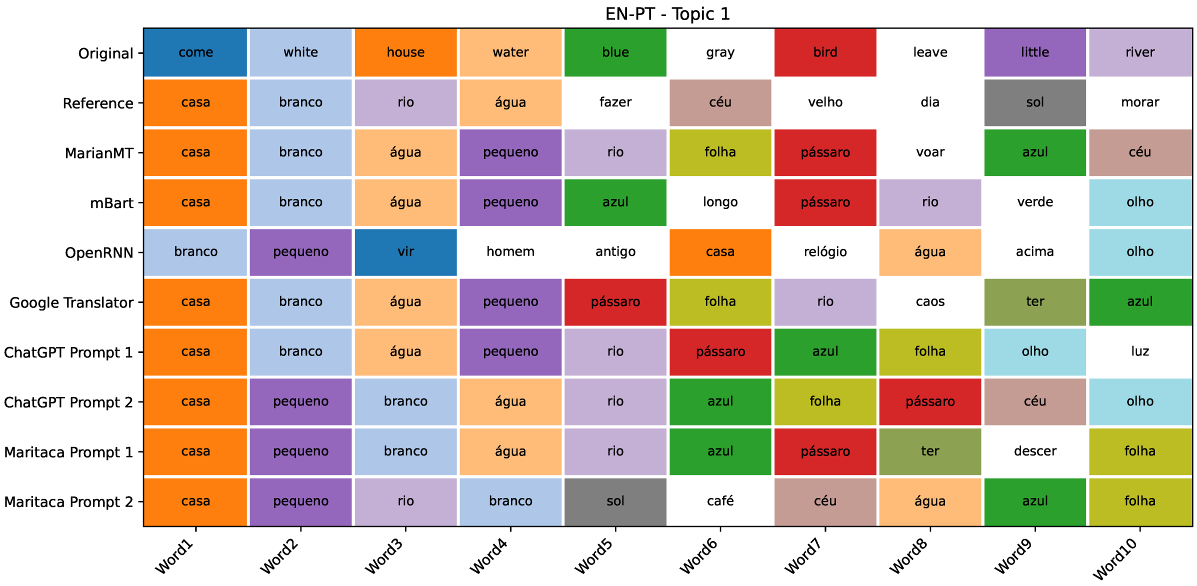


	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	love	time	lose	say	live	fall	bye	sweet	eye	pain
Reference	amour	perdre	vouloir	pitié	jeune	oeil	temps	adieu	jour	souhaiter
MarianMT	amour	faire	perdre	temps	douleur	tomber	revoir	manquer	mal	vie
mBart	vie	amour	guilde	perdre	douleur	oeil	ambition	jeune	fin	souffle
OpenRNN	amour	douleur	fin	pâle	recherche	oeil	noir	esprit	perdre	bon
Google Translator	amour	temps	entier	jeune	perdre	vie	fille	ambition	atteindre	mourir
ChatGPT Prompt 1	amour	temps	fille	jeune	perdre	douleur	ambition	tomber	peine	pitié
ChatGPT Prompt 2	amour	temps	revoir	perdre	ambition	pitié	faire	malade	grâce	dévorer
Maritaca Prompt 1	amour	temps	vouloir	perdre	jeune	laisser	douleur	ambition	fille	jouer
Maritaca Prompt 2	amour	temps	effroi	jeune	perdre	oeil	fille	saveur	mesure	sortir

Source: Own authorship.

English-Portuguese

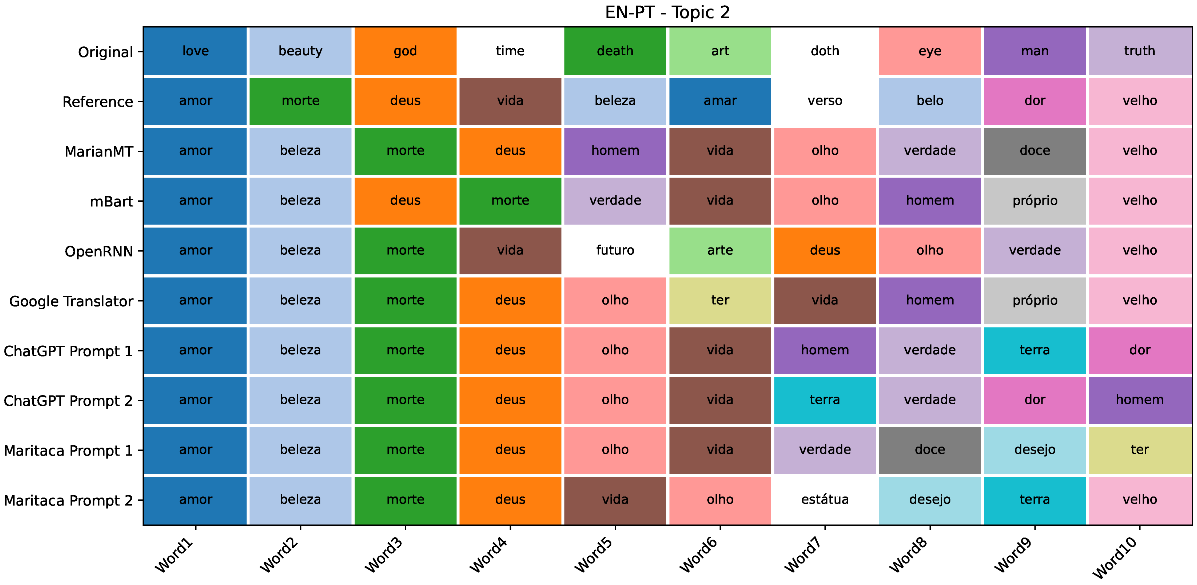
Figure 20 – Keywords for Topic 1 in the original poem and its translations for the English-Portuguese language pair.



	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	come	white	house	water	blue	gray	bird	leave	little	river
Reference	casa	branco	rio	água	fazer	céu	velho	dia	sol	morar
MarianMT	casa	branco	água	pequeno	rio	folha	pássaro	voar	azul	céu
mBart	casa	branco	água	pequeno	azul	longo	pássaro	rio	verde	olho
OpenRNN	branco	pequeno	vir	homem	antigo	casa	relógio	água	acima	olho
Google Translator	casa	branco	água	pequeno	pássaro	folha	rio	caos	ter	azul
ChatGPT Prompt 1	casa	branco	água	pequeno	rio	pássaro	azul	folha	olho	luz
ChatGPT Prompt 2	casa	pequeno	branco	água	rio	azul	folha	pássaro	céu	olho
Maritaca Prompt 1	casa	pequeno	branco	água	rio	azul	pássaro	ter	descer	folha
Maritaca Prompt 2	casa	pequeno	rio	branco	sol	café	céu	água	azul	folha

Source: Own authorship.

Figure 21 – Keywords for Topic 2 in the original poem and its translations for the English-Portuguese language pair.

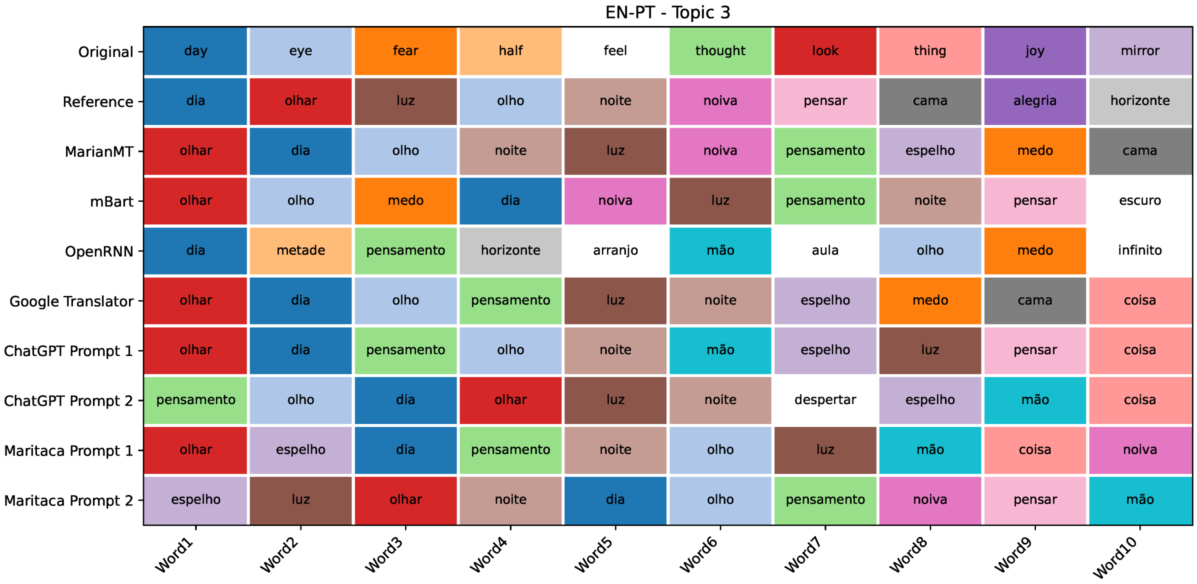


EN-PT - Topic 2

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	love	beauty	god	time	death	art	doth	eye	man	truth
Reference	amor	morte	deus	vida	beleza	amar	verso	belo	dor	velho
MarianMT	amor	beleza	morte	deus	homem	vida	olho	verdade	doce	velho
mBart	amor	beleza	deus	morte	verdade	vida	olho	homem	próprio	velho
OpenRNN	amor	beleza	morte	vida	futuro	arte	deus	olho	verdade	velho
Google Translator	amor	beleza	morte	deus	olho	ter	vida	homem	próprio	velho
ChatGPT Prompt 1	amor	beleza	morte	deus	olho	vida	homem	verdade	terra	dor
ChatGPT Prompt 2	amor	beleza	morte	deus	olho	vida	terra	verdade	dor	homem
Maritaca Prompt 1	amor	beleza	morte	deus	olho	vida	verdade	doce	desejo	ter
Maritaca Prompt 2	amor	beleza	morte	deus	vida	olho	estátua	desejo	terra	velho

Source: Own authorship.

Figure 22 – Keywords for Topic 3 in the original poem and its translations for the English-Portuguese language pair.



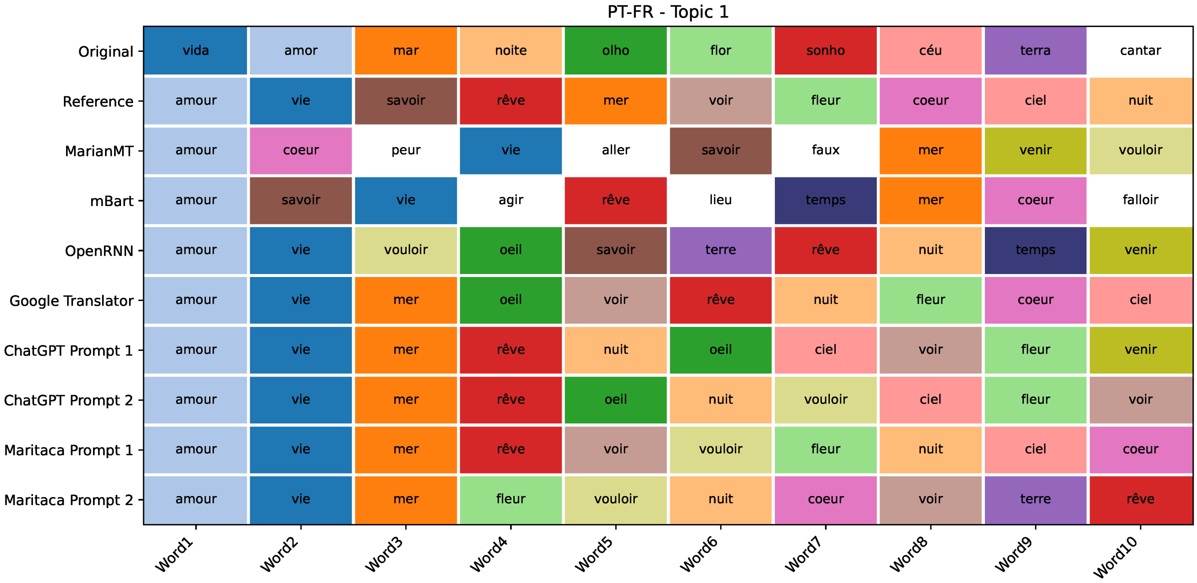
EN-PT - Topic 3

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	day	eye	fear	half	feel	thought	look	thing	joy	mirror
Reference	dia	olhar	luz	olho	noite	noiva	pensar	cama	alegria	horizonte
MarianMT	olhar	dia	olho	noite	luz	noiva	pensamento	espelho	medo	cama
mBart	olhar	olho	medo	dia	noiva	luz	pensamento	noite	pensar	escuro
OpenRNN	dia	metade	pensamento	horizonte	arranjo	mão	aula	olho	medo	infinito
Google Translator	olhar	dia	olho	pensamento	luz	noite	espelho	medo	cama	coisa
ChatGPT Prompt 1	olhar	dia	pensamento	olho	noite	mão	espelho	luz	pensar	coisa
ChatGPT Prompt 2	pensamento	olho	dia	olhar	luz	noite	despertar	espelho	mão	coisa
Maritaca Prompt 1	olhar	espelho	dia	pensamento	noite	olho	luz	mão	coisa	noiva
Maritaca Prompt 2	espelho	luz	olhar	noite	dia	olho	pensamento	noiva	pensar	mão

Source: Own authorship.

Portuguese-French

Figure 23 – Keywords for Topic 1 in the original poem and its translations for the Portuguese-French language pair.

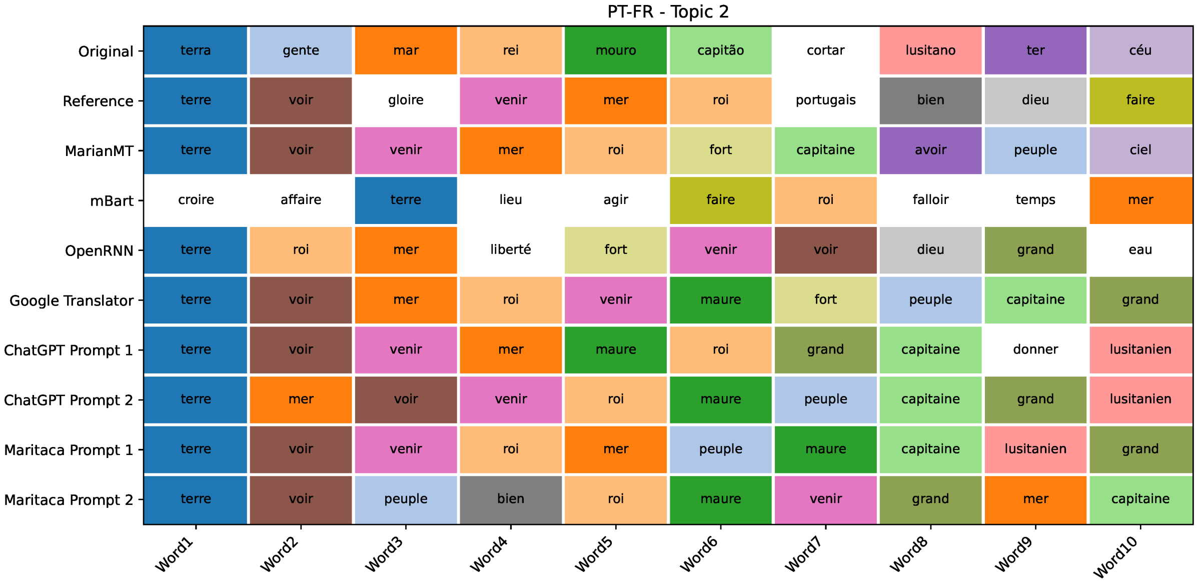


PT-FR - Topic 1

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	vida	amor	mar	noite	olho	flor	sonho	céu	terra	cantar
Reference	amour	vie	savoir	rêve	mer	voir	fleur	coeur	ciel	nuit
MarianMT	amour	coeur	peur	vie	aller	savoir	faux	mer	venir	vouloir
mBart	amour	savoir	vie	agir	rêve	lieu	temps	mer	coeur	falloir
OpenRNN	amour	vie	vouloir	oeil	savoir	terre	rêve	nuit	temps	venir
Google Translator	amour	vie	mer	oeil	voir	rêve	nuit	fleur	coeur	ciel
ChatGPT Prompt 1	amour	vie	mer	rêve	nuit	oeil	ciel	voir	fleur	venir
ChatGPT Prompt 2	amour	vie	mer	rêve	oeil	nuit	vouloir	ciel	fleur	voir
Maritaca Prompt 1	amour	vie	mer	rêve	voir	vouloir	fleur	nuit	ciel	coeur
Maritaca Prompt 2	amour	vie	mer	fleur	vouloir	nuit	coeur	voir	terre	rêve

Source: Own authorship.

Figure 24 – Keywords for Topic 2 in the original poem and its translations for the Portuguese-French language pair.

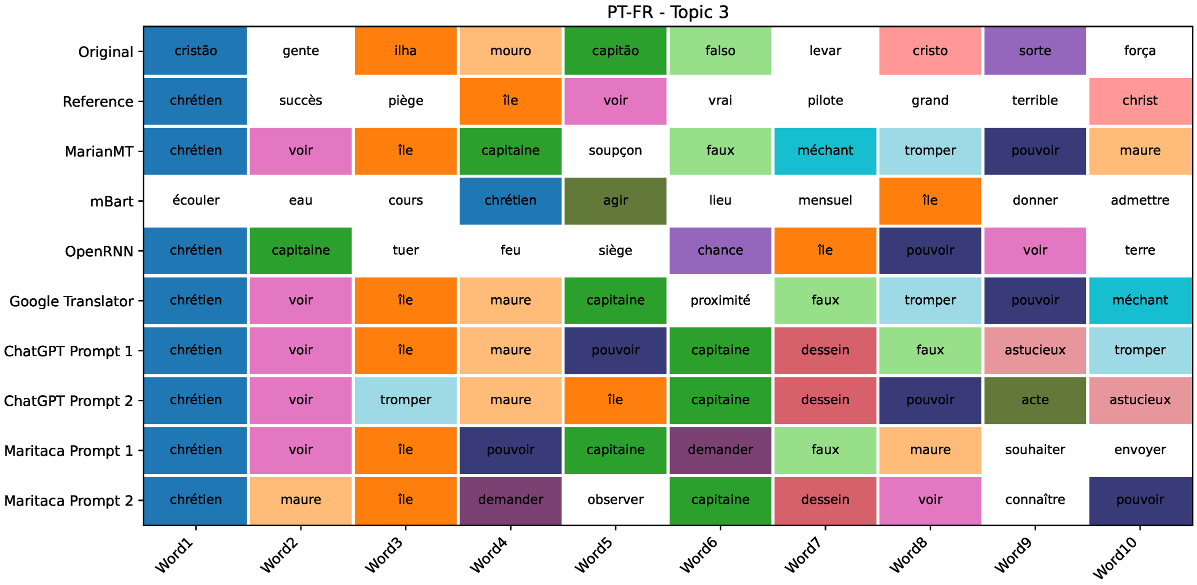


PT-FR - Topic 2

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	terra	gente	mar	rei	mouro	capitão	cortar	lusitano	ter	céu
Reference	terre	voir	gloire	venir	mer	roi	portugais	bien	dieu	faire
MarianMT	terre	voir	venir	mer	roi	fort	capitaine	avoir	peuple	ciel
mBart	croire	affaire	terre	lieu	agir	faire	roi	falloir	temps	mer
OpenRNN	terre	roi	mer	liberté	fort	venir	voir	dieu	grand	eau
Google Translator	terre	voir	mer	roi	venir	maure	fort	peuple	capitaine	grand
ChatGPT Prompt 1	terre	voir	venir	mer	maure	roi	grand	capitaine	donner	lusitanien
ChatGPT Prompt 2	terre	mer	voir	venir	roi	maure	peuple	capitaine	grand	lusitanien
Maritaca Prompt 1	terre	voir	venir	roi	mer	peuple	maure	capitaine	lusitanien	grand
Maritaca Prompt 2	terre	voir	peuple	bien	roi	maure	venir	grand	mer	capitaine

Source: Own authorship.

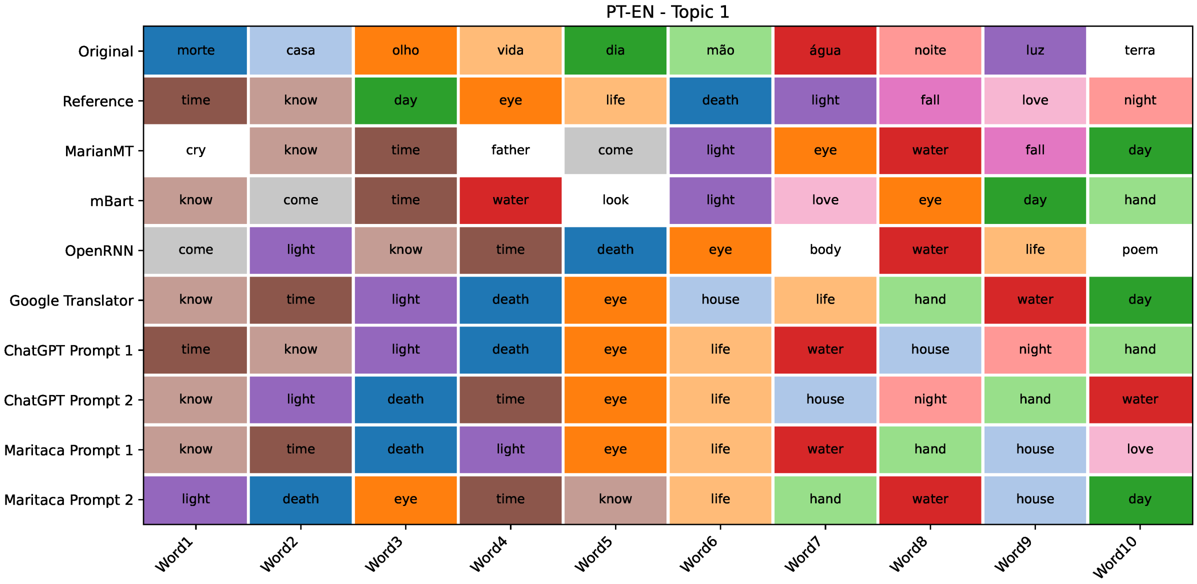
Figure 25 – Keywords for Topic 3 in the original poem and its translations for the Portuguese-French language pair.



Source: Own authorship.

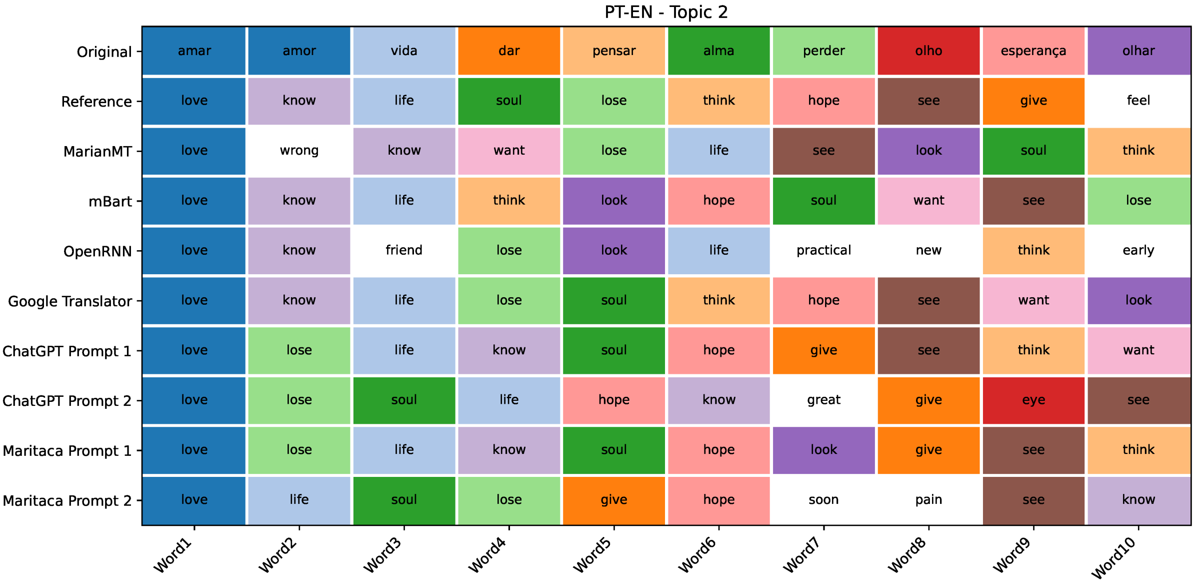
Portuguese-English

Figure 26 – Keywords for Topic 1 in the original poem and its translations for the Portuguese-English language pair.



Source: Own authorship.

Figure 27 – Keywords for Topic 2 in the original poem and its translations for the Portuguese-English language pair.



PT-EN - Topic 2

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Original	amar	amor	vida	dar	pensar	alma	perder	olho	esperança	olhar
Reference	love	know	life	soul	lose	think	hope	see	give	feel
MarianMT	love	wrong	know	want	lose	life	see	look	soul	think
mBart	love	know	life	think	look	hope	soul	want	see	lose
OpenRNN	love	know	friend	lose	look	life	practical	new	think	early
Google Translator	love	know	life	lose	soul	think	hope	see	want	look
ChatGPT Prompt 1	love	lose	life	know	soul	hope	give	see	think	want
ChatGPT Prompt 2	love	lose	soul	life	hope	know	great	give	eye	see
Maritaca Prompt 1	love	lose	life	know	soul	hope	look	give	see	think
Maritaca Prompt 2	love	life	soul	lose	give	hope	soon	pain	see	know

Source: Own authorship.

APPENDIX B

Poems and Tables for the Human Evaluation

This appendix provides the materials and results related to the human evaluation. Section B.1 presents the original poems and their translations used in the assessment, while Section B.2 includes the complete score tables assigned by each evaluator to the automatic translations across all language pairs.

B.1 Poems and Translations Used in the Human Evaluation

French-English

Clair de lune **(Paul Verlaine)**

Original (French):

Votre âme est un paysage choisi
Que vont charmant masques et bergamasques
Jouant du luth et dansant et quasi
Tristes sous leurs déguisements fantasques.

Tout en chantant sur le mode mineur
L'amour vainqueur et la vie opportune
Ils n'ont pas l'air de croire à leur bonheur
Et leur chanson se mêle au clair de lune,

Au calme clair de lune triste et beau,
Qui fait rêver les oiseaux dans les arbres
Et sangloter d'extase les jets d'eau,
Les grands jets d'eau sveltes parmi les marbres.

Human Reference (Norman R. Shapiro):

Your soul is a chosen landscape
On which masks and Bergamasques cast enchantment as they go,
Playing the lute, and dancing, and all but
Sad beneath their fantasy-disguises.

Singing all the while, in the minor mode,
Of all-conquering love and life so kind to them
They do not seem to believe in their good fortune,
And their song mingles with the moonlight,

With the calm moonlight, sad and lovely,
Which makes the birds dream in the trees,
And the plumes of the fountains weep in ecstasy,
The tall, slender plumes of the fountains among the marble sculptures.

Specialized MT (Google Translator):

Your soul is a chosen landscape
Which charming masks and Bergamasques go
Playing the lute and dancing and almost
Sad beneath their fanciful disguises.

While singing in a minor key
Victorious love and opportune life
They don't seem to believe in their happiness
And their song mingles with the moonlight,

With the calm, sad and beautiful moonlight,
Which makes the birds dream in the trees
And the fountains sob with ecstasy,
The great, slender fountains among the marble.

LLM (ChatGPT Prompt 1):

Your soul is a chosen landscape
Where charming masks and bergamasks pass by,
Playing the lute and dancing, almost
Sad beneath their whimsical disguises.

Even while singing in a minor key
Of conquering love and life's fair chances,

They seem not to believe in their own joy,
And their song mingles with the moonlight,

With the still, clear moonlight, sad and lovely,
That sets the birds dreaming in the trees
And makes the fountains sob with ecstasy–
Tall, slender fountains among the marble.

French-Portuguese

Demain, dès l'aube (Victor Hugo)

Original (French):

Demain, dès l'aube, à l'heure où blanchit la campagne,
Je partirai. Vois-tu, je sais que tu m'attends.
J'irai par la forêt, j'irai par la montagne.
Je ne puis demeurer loin de toi plus longtemps.

Je marcherai les yeux fixés sur mes pensées,
Sans rien voir au dehors, sans entendre aucun bruit,
Seul, inconnu, le dos courbé, les mains croisées,
Triste, et le jour pour moi sera comme la nuit.

Je ne regarderai ni l'or du soir qui tombe,
Ni les voiles au loin descendant vers Harfleur,
Et quand j'arriverai, je mettrai sur ta tombe
Un bouquet de houx vert et de bruyère en fleur.

Human Reference (JB Xavier):

Amanhã, ao amanhecer, quando de branco o campo se banha,
Eu partirei. Sei que você me vê, e sei que espera por mim...
Seguirei pelas florestas e cruzarei a montanha.
Não posso ficar longe de você tanto tempo assim...

Marcharei com os olhos fixos em meu pensamento,
Sem desviar minha atenção, sem ouvir qualquer ruído,
Só e desconhecido, costas encurvadas, as mãos num lamento,
Dia triste como a noite para mim, quando eu tiver partido.

Nem para o dourado da noite que cai eu olharei,
Nem verei as velas que para Harfleur descem em triste ardor...
E quando eu chegar, em seu túmulo colocarei

Um buquê de verde azevinho e urzes em flor.

Specialized MT (Google Translator):

Amanhã, ao amanhecer, na hora em que o campo embranquecer,
Partirei. Sabe, sei que está me esperando.
Atravessarei a floresta, atravessarei as montanhas.
Não posso mais ficar longe de você.

Caminharei com os olhos fixos nos meus pensamentos,
Sem ver nada lá fora, sem ouvir nenhum som,
Sozinha, desconhecida, com as costas curvadas, as mãos cruzadas,
Triste, e o dia será como a noite para mim.

Não olharei para o dourado da tarde caindo,
Nem para as velas distantes descendo em direção a Harfleur,
E quando eu chegar, colocarei em seu túmulo
Um buquê de azevinho verde e urze em flor.

LLM (Maritaca Prompt 2):

Amanhã, ao romper da aurora, quando a campina se alveja,
Eu partirei. Sabes, eu sei que me aguardas.
Irei pela floresta, irei pela montanha.
Não posso permanecer longe de ti por mais tempo.

Caminharei com os olhos fixos em meus pensamentos,
Sem nada ver ao redor, sem ouvir qualquer rumor,
Só, incógnito, o dorso curvado, as mãos cruzadas,
Triste, e o dia para mim será como a noite.

Não olharei o ouro do crepúsculo que desce,
Nem as velas ao longe que rumam para Harfleur,
E quando eu chegar, sobre tua sepultura,
Um ramo de azevinho verde e urze em flor eu colocarei.

English-French

The Road Not Taken **(Robert Frost)**

Original (English):

Two roads diverged in a yellow wood,
And sorry I could not travel both

And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I –
I took the one less traveled by,
And that has made all the difference.

Human Reference (François Charton):

Dans un bois jaune deux chemins se séparaient
Et désolé de ne pouvoir emprunter l'un et l'autre
Et n'être qu'un voyageur longtemps je demeurai
Et suivis du regard autant que je pouvais
Le premier jusqu'au tournant sous les hêtres ;

Puis je pris le second, presqu'aussi tentant
Et peut-être même plus recommandé
Parce que tout herbeux, à la marche invitant ;
Même si, pour dire vrai les passants
Les avaient tous deux pareillement usés

L'un et l'autre s'offraient au petit jour
Avec leurs feuilles qu'aucun marcheur n'avait noircies
Je gardais le premier pour un autre parcours
Tout en sachant que de tour en détour
Je ne repasserais probablement jamais ici

C'est en soupirant que je devrais l'avouer
Je ne sais où, il y a bien longtemps

Deux chemins dans un bois se séparaient et j'ai –
J'ai pris des deux le moins fréquenté
Et c'était sans doute le plus important.

Specialized MT (Google Translator):

Deux routes divergeaient dans un bois jaune,
Et désolé de ne pouvoir les parcourir toutes les deux
Et être un seul voyageur, je suis resté longtemps
Et j'ai regardé l'une aussi loin que possible
Jusqu'à l'endroit où elle serpentait dans les sous-bois ;

Puis j'ai pris l'autre, tout aussi belle
Et ayant peut-être le meilleur titre,
Parce qu'elle était herbeuse et avait besoin d'être usée ;
Bien que le passage là-bas
Les ait vraiment usées autant,

Et toutes deux ce matin-là, également
Dans des feuilles qu'aucun pas n'avait foulées noires.
Oh, j'ai gardé la première pour un autre jour !
Mais sachant comment un chemin mène à un autre chemin,
je doutais de revenir un jour.

Je raconterai cela avec un soupir
Quelque part dans des siècles et des siècles :
Deux routes divergeaient dans un bois, et je…
J'ai pris la moins fréquentée,
Et cela a fait toute la différence.

LLM (ChatGPT Prompt 1):

Deux routes divergeaient dans un bois jaunissant,
Et, désolé de ne pouvoir les suivre toutes deux,
Étant un seul voyageur, longtemps je me tins là,
Regardant l'une aussi loin que possible,
Jusqu'à ce qu'elle se courbe dans les broussailles ;

Puis je pris l'autre, tout aussi belle,
Et qui semblait peut-être avoir meilleur titre,
Parce qu'elle était herbeuse et peu foulée ;
Quoique, en réalité, le passage
Les eût usées à peu près également.

Et ce matin-là, elles gisaient toutes deux
Sous des feuilles que nul pas n'avait noircies.
Oh, j'ai gardé la première pour un autre jour !
Mais sachant comment un chemin en mène à un autre,
Je doutais de jamais revenir sur mes pas.

Je raconterai cela avec un soupir
Quelque part, dans bien des années :
Deux routes divergeaient dans un bois, et moi –
J'ai pris celle que l'on avait moins suivie,
Et cela a fait toute la différence.

English-Portuguese

Do not go gentle into that good night (Dylan Thomas)

Original (English):

Do not go gentle into that good night,
Old age should burn and rave at close of day;
Rage, rage against the dying of the light.

Though wise men at their end know dark is right,
Because their words had forked no lightning they
Do not go gentle into that good night.

Good men, the last wave by, crying how bright
Their frail deeds might have danced in a green bay,
Rage, rage against the dying of the light.

Wild men who caught and sang the sun in flight,
And learn, too late, they grieved it on its way,
Do not go gentle into that good night.

Grave men, near death, who see with blinding sight
Blind eyes could blaze like meteors and be gay,
Rage, rage against the dying of the light.

And you, my father, there on the sad height,
Curse, bless, me now with your fierce tears, I pray.
Do not go gentle into that good night.
Rage, rage against the dying of the light.

Human Reference (Anderson Braga Horta):

Não entres mansamente nessa noite funda.
Que as velhas almas ardem ao findar do dia.
Te insurge em fúria contra o fim da luz, e luta.

Os sábios, mesmo vendo a sombra que triunfa,
Sabendo que sua voz não fulge nem fulmina,
Não entram mansamente nessa noite funda.

Os bravos, ao romper das ondas, não se assustam,
Mas cantam suas proezas na enseada limpa:
Em fúria rugem contra o fim da luz, e lutam.

Os bárbaros, que aos brados catam sol e lua,
E, súbito, lamentam que essa luz se extinga,
Não entram mansamente nessa noite funda.

Os quase mortos – cegos, lúcidos – perscrutam
Clarões dos meteoros cegos da alegria:
Em fúria rugem contra o fim da luz, e lutam.

E tu, meu pai, erguido em tormentosa altura,
Com lágrima feroz me amaldiçoa e guia.
Não entres mansamente nessa noite funda;
Te insurge em fúria contra o fim da luz, e lutam.

Specialized MT (Google Translator):

Não entrem mansos nessa boa noite,
A velhice deve queimar e delirar ao fim do dia;
Fúria, fúria contra o morrer da luz.

Embora os sábios, em seu fim, saibam que a escuridão é certa,
Porque suas palavras não bifurcaram nenhum relâmpago,
Não entrem mansos nessa boa noite.

Bons homens, a última onda que passou, gritando o quão brilhante
Seus atos frágeis poderiam ter dançado em uma baía verde,
Fúria, fúria contra o morrer da luz.

Homens selvagens que capturaram e cantaram o sol em voo,
E aprenderam, tarde demais, que o entristeceram em seu caminho,
Não entrem mansos nessa boa noite.

Homens graves, à beira da morte, que enxergam com uma visão ofuscante

Olhos cegos poderiam brilhar como meteoros e ser alegres,
Fúria, fúria contra o morrer da luz.

E você, meu pai, lá no alto triste,
Amaldiçoe-me, abençoe-me agora com suas lágrimas ferozes, eu oro.
Não entrem mansos nessa boa noite.
Fúria, fúria contra o morrer da luz.

LLM (Maritaca Prompt 2):

Não entres docilmente nessa noite serena,
Que a velhice arda e brade ao fim do dia;
Brada, brada contra o apagar da luz.

Embora sábios saibam, ao fim, que a noite é certa,
Porque suas palavras não roubaram o raio,
Não entres docilmente nessa noite serena.

Homens bons, na última onda, clamam quão brilhante
Suas frágeis ações poderiam ter dançado em um verde golfo,
Brada, brada contra o apagar da luz.

Homens selvagens que capturaram e cantaram o sol em fuga,
E aprendem, tarde demais, que o lamentaram em sua jornada,
Não entres docilmente nessa noite serena.

Homens graves, ao morrer, que veem com visão que cega
Olhos cegos poderiam brilhar como meteoros e ser alegres,
Brada, brada contra o apagar da luz.

E tu, meu pai, lá na triste elevação,
Amaldiçoa, abençoa-me agora com tuas ferozes lágrimas, eu rogo.
Não entres docilmente nessa noite serena.
Brada, brada contra o apagar da luz.

Portuguese-French

Traduzir-se **(Ferreira Gullar)**

Original (Portuguese):

Uma parte de mim
é todo mundo:
outra parte é ninguém:
fundo sem fundo.

Uma parte de mim
é multidão:
outra parte estranheza
e solidão.

Uma parte de mim
pesa, pondera:
outra parte
delira.

Uma parte de mim
almoça e janta:
outra parte
se espanta.

Uma parte de mim
é permanente:
outra parte
se sabe de repente.

Uma parte de mim
é só vertigem:
outra parte,
linguagem.

Traduzir uma parte
na outra parte
– que é uma questão
de vida ou morte –
será arte?

Human Reference (L. Gonçalves et D. Lamaison):

Une part de moi
est tout le monde :
l'autre part est personne
fond que rien ne fonde.

Une part de moi
est multitude :
et l'autre étrangeté
et solitude.

Une part de moi

pèse, médite :
l'autre part
se délite.

Une part de moi
déjeune et dîne :
l'autre part
halluciné.

Une part de moi
toujours se connaît :
L'autre part
tout soudain se sait.

Une part de moi
n'est que tangage :
l'autre part,
langage.

Traduire une part
dans l'autre part
- ce qui est une question
de vie ou mort -
est-ce l'art ?

Specialized MT (Google Translator):

Une partie de moi
est tout le monde :
une autre partie n'est personne :
une profondeur sans fond.

Une partie de moi
est multitude :
une autre partie étrangeté
et solitude.

Une partie de moi
pese, médite :
une autre partie
s'extasie.

Une partie de moi
mange et dîne :

une autre partie
s’étonne.

Une partie de moi
est permanente :
une autre partie
sait soudain.

Une partie de moi
n’est que vertige:
une autre partie,
le langage.

Traduire une partie
en une autre partie
– ce qui est une question de vie ou de mort –
serait-ce de l’art ?

LLM (ChatGPT Prompt 1):

Une partie de moi
est tout le monde :
une autre partie n’est personne :
profondeur sans fond.

Une partie de moi
est une foule :
une autre partie, étrangeté
et solitude.

Une partie de moi
pèse, réfléchit :
une autre partie
délire.

Une partie de moi
déjeune et dîne :
une autre partie
s’émerveille.

Une partie de moi
est permanente :
une autre partie
se découvre soudainement.

Une partie de moi
n’est que vertige :
une autre partie,
langage.

Traduire une partie
dans l’autre partie
– ce qui est une question
de vie ou de mort –
serait-ce de l’art ?

Portuguese-English

Canção do Exílio **(Gonçalves Dias)**

Original (Portuguese):

Minha terra tem palmeiras,
Onde canta o Sabiá;
As aves, que aqui gorjeiam,
Não gorjeiam como lá.

Nosso céu tem mais estrelas,
Nossas várzeas têm mais flores,
Nossos bosques têm mais vida,
Nossa vida mais amores.

Em cismar, sozinho, à noite,
Mais prazer eu encontro lá;
Minha terra tem palmeiras,
Onde canta o Sabiá.

Minha terra tem primores,
Que tais não encontro eu cá;
Em cismar –sozinho, à noite–
Mais prazer eu encontro lá;
Minha terra tem palmeiras,
Onde canta o Sabiá.

Não permita Deus que eu morra,
Sem que eu volte para lá;
Sem que disfrute os primores
Que não encontro por cá;
Sem qu’inda aviste as palmeiras,

Onde canta o Sabiá.

Human Reference (Nelson Ascher):

My homeland has many palm-trees
and the thrush-song fills its air;
no bird here can sing as well
as the birds sing over there

We have fields more full of flowers
and a starrier sky above,
we have woods more full of life
and a life more full of love.

Lonely night-time meditations
please me more when I am there;
my homeland has many palm-trees
and the thrush-song fills its air.

Such delights as my land offers
Are not found here nor elsewhere;
lonely night-time meditations
please me more when I am there;
My homeland has many palm-trees
and the thrush-song fills its air.

Don't allow me, God, to die
without getting back to where
I belong, without enjoying
the delights found only there,
without seeing all those palm-trees,
hearing thrush-songs fill the air.

Specialized MT (Google Translator):

My land has palm trees,
Where the thrush sings;
The birds that chirp here,
Don't chirp like they do there.

Our sky has more stars,
Our meadows have more flowers,
Our forests have more life,

Our life more loves.

In brooding alone at night,
I find more pleasure there;
My land has palm trees,
Where the thrush sings.

My land has delights,
I don't find such here;
In brooding–alone at night–
I find more pleasure there;
My land has palm trees,
Where the thrush sings.

God forbid that I die,
Without returning there;
Without enjoying the delights
I don't find here;
Without still seeing the palm trees,
Where the thrush sings.

LLM (ChatGPT Prompt 1):

My land has palm trees,
Where the thrush sings;
The birds that sing here
Do not sing as they do there.

Our sky has more stars,
Our lowlands have more flowers,
Our forests have more life,
Our life more loves.

In musing, alone at night,
More pleasure I find there;
My land has palm trees,
Where the thrush sings.

My land holds wonders
That I do not find here;
In musing – alone, at night –
More pleasure I find there;
My land has palm trees,
Where the thrush sings.

May God not allow me to die
Without returning there;
Without enjoying the wonders
I do not find here;
Without still beholding the palm trees,
Where the thrush sings.

B.2 Complete Tables of Human Evaluation Scores

French-English

Table 18 – Human evaluation scores for the specialized MT system (Google Translator) for the French-English pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	1	3	3	2	2.2
2	1	2	3	3	2	2.2
3	2	2	2	3	2	2.2
4	1	1	3	2	2	1.8
5	2	1	3	3	2	2.2
Average per Criterion	1.6	1.4	2.8	2.8	2.0	
Standard Deviation	0.49	0.49	0.45	0.45	0.0	

Source: Own authorship.

Table 19 – Human evaluation scores for the LLM (ChatGPT Prompt 1) for the French-English pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	2	4	4	3	3.0
2	2	1	4	4	3	2.8
3	2	2	3	4	3	2.8
4	1	2	4	3	3	2.6
5	2	2	4	4	4	3.2
Average per Criterion	1.8	1.8	3.8	3.8	3.2	
Standard Deviation	0.42	0.42	0.45	0.45	0.37	

Source: Own authorship.

French-Portuguese

Table 20 – Human evaluation scores for the specialized MT system (Google Translator) for the French-Portuguese pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	1	3	3	2	2.2
2	1	2	3	3	2	2.2
3	2	2	3	3	2	2.4
4	1	1	2	3	2	1.8
5	2	1	3	3	3	2.4
Average per Criterion	1.6	1.4	2.8	3.0	2.2	
Standard Deviation	0.49	0.49	0.45	0.0	0.45	

Source: Own authorship.

Table 21 – Human evaluation scores for the LLM (Maritaca Prompt 2) for the French-Portuguese pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	2	4	4	3	3.0
2	2	1	4	4	3	2.8
3	2	2	3	4	3	2.8
4	1	2	4	3	3	2.6
5	2	2	4	4	4	3.2
Average per Criterion	1.8	1.8	3.8	3.8	3.2	
Standard Deviation	0.42	0.42	0.45	0.45	0.37	

Source: Own authorship.

English-French

Table 22 – Human evaluation scores for the specialized MT system (Google Translator) for the English-French pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	1	3	3	2	2.2
2	1	2	3	3	2	2.2
3	2	2	3	3	2	2.4
4	1	1	2	3	2	2.0
5	2	1	3	3	3	2.4
Average per Criterion	1.6	1.4	2.8	3.0	2.2	
Standard Deviation	0.49	0.49	0.45	0.0	0.45	

Source: Own authorship.

Table 23 – Human evaluation scores for the LLM (ChatGPT Prompt 1) for the English-French pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	2	4	4	3	3.0
2	2	1	4	4	3	2.8
3	2	2	4	4	4	3.2
4	1	2	4	3	3	2.8
5	2	2	4	4	4	3.2
Average per Criterion	1.8	1.8	4.0	3.8	3.4	
Standard Deviation	0.42	0.42	0.0	0.45	0.37	

Source: Own authorship.

English-Portuguese

Table 24 – Human evaluation scores for the specialized MT system (Google Translator) for the English-Portuguese pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	1	3	3	2	2.2
2	1	2	3	3	2	2.2
3	2	1	3	3	3	2.4
4	1	1	2	3	2	1.8
5	2	2	3	3	2	2.4
Average per Criterion	1.6	1.4	2.8	3.0	2.2	
Standard Deviation	0.49	0.49	0.45	0.0	0.45	

Source: Own authorship.

Table 25 – Human evaluation scores for the LLM (Maritaca Prompt 2) for the English-Portuguese pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	2	4	4	3	3.0
2	2	1	4	4	3	2.8
3	2	2	4	4	4	3.2
4	1	2	4	3	3	2.8
5	2	2	4	4	4	3.2
Average per Criterion	1.8	1.8	4.0	3.8	3.4	
Standard Deviation	0.42	0.42	0.0	0.45	0.37	

Source: Own authorship.

Portuguese-French

Table 26 – Human evaluation scores for the specialized MT system (Google Translator) for the Portuguese-French pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	1	3	3	2	2.2
2	1	2	3	3	2	2.2
3	2	1	3	3	3	2.4
4	1	1	2	3	2	1.8
5	2	2	3	3	2	2.4
Average per Criterion	1.6	1.4	2.8	3.0	2.2	
Standard Deviation	0.49	0.49	0.45	0.0	0.45	

Source: Own authorship.

Table 27 – Human evaluation scores for the LLM (ChatGPT Prompt 1) for the Portuguese-French pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	2	4	4	3	3.0
2	2	1	4	4	3	2.8
3	2	2	4	4	4	3.2
4	1	2	4	3	3	2.8
5	2	2	4	4	4	3.2
Average per Criterion	1.8	1.8	4.0	3.8	3.4	
Standard Deviation	0.42	0.42	0.0	0.45	0.37	

Source: Own authorship.

Portuguese-English

Table 28 – Human evaluation scores for the specialized MT system (Google Translator) for the Portuguese-English pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	1	3	3	2	2.2
2	2	2	3	3	2	2.4
3	1	2	3	3	2	2.2
4	2	1	2	3	2	2.0
5	1	2	3	3	3	2.4
Average per Criterion	1.6	1.6	2.8	3.0	2.2	
Standard Deviation	0.49	0.49	0.45	0.0	0.45	

Source: Own authorship.

Table 29 – Human evaluation scores for the LLM (ChatGPT Prompt 1) for the Portuguese-English pair.

Specialist	**Poetic Structure**	**Poetic Stylistics**	**Fluency**	**Meaning**	**Overall Impression**	**Average**
1	2	2	4	4	3	3.0
2	2	2	4	4	3	3.0
3	2	1	4	4	4	3.0
4	1	2	4	4	3	2.8
5	2	2	4	4	4	3.2
Average per Criterion	1.8	1.8	4.0	4.0	3.4	
Standard Deviation	0.42	0.42	0.0	0.0	0.45	

Source: Own authorship.

APPENDIX C

Results of Module 2 - Lyrics-Adapted (BLEU and METEOR)

This appendix presents the detailed BLEU and METEOR results from Module 2, which examined the impact of fine-tuning on song lyrics for each specialized translation model. These results complement the BERTScore analyses discussed in the main text. Table 30 compares the fine-tuned and pre-trained versions of mBART, MarianMT, and OpenNMT (RNN) across the six language pairs. The bold values indicate the best results among the fine-tuning strategies for each translation model and language pair.

Table 30 – Evaluation results for all MT models (Lyrics Fine-Tuning vs. Pre-Trained) using BLEU and METEOR.

Model	**Metric**	**Language Pair**	**Lyrics Fine-Tuning**	**Pre-Trained**
mBART	BLEU	FR-EN	0.1638	**0.2751**
		FR-PT	**0.1761**	0.0628
		EN-FR	0.0770	**0.1421**
		EN-PT	0.1045	**0.1163**
		PT-FR	0.0628	**0.0817**
		PT-EN	**0.1743**	0.1403
	METEOR	FR-EN	0.3348	**0.5462**
		FR-PT	**0.4314**	0.2339
		EN-FR	0.2523	**0.3886**
		EN-PT	0.3038	**0.3781**
		PT-FR	0.1837	**0.2562**
		PT-EN	0.3616	**0.4202**
MarianMT	BLEU	FR-EN	0.1527	**0.2948**
		FR-PT	0.0070	**0.1650**
		EN-FR	0.1309	**0.1805**
		EN-PT	0.0169	**0.1480**
		PT-FR	0.0822	**0.1899**
		PT-EN	0.0928	**0.2411**
	METEOR	FR-EN	0.4437	**0.5755**
		FR-PT	0.1084	**0.4241**
		EN-FR	0.3782	**0.4369**
		EN-PT	0.1729	**0.4202**
		PT-FR	0.3025	**0.4049**
		PT-EN	0.3737	**0.5213**
OpenNMT	BLEU	FR-EN	0.0102	**0.0526**
		FR-PT	0.0044	**0.0303**
		EN-FR	0.0041	**0.0226**
		EN-PT	0.0033	**0.0143**
		PT-FR	0.0066	**0.0586**
		PT-EN	0.0079	**0.0546**
	METEOR	FR-EN	0.0891	**0.2632**
		FR-PT	0.0665	**0.2062**
		EN-FR	0.0638	**0.1911**
		EN-PT	0.0712	**0.1934**
		PT-FR	0.0796	**0.2337**
		PT-EN	0.0880	**0.2991**

Source: Own authorship.

APPENDIX D

Results of Module 3 - Poems and Lyrics-Adapted (BERTScore (LLMs), BLEU and METEOR)

This appendix presents the quantitative results from Module 3, which compared fine-tuning strategies using poems and a combination of poems and lyrics. The analyses include BLEU and METEOR scores for specialized machine translation models – mBART, MarianMT, and OpenNMT (RNN) – as well as the BERTScore, BLEU, and METEOR results for LLMs.

Table 31 summarizes the performance of specialized MT models under different fine-tuning configurations. Table 32 highlights the best-performing scores across these models, including Google Translate as an additional baseline. Table 33 reports the results for ChatGPT and Maritaca AI under two prompting strategies, using the same test set as the MT systems. The bold values indicate the best results among the fine-tuning strategies for each translation model and language pair.

Table 31 – Automatic evaluation metrics (BLEU, METEOR) for mBART, MarianMT, and OpenNMT (RNN) translations with different fine-tuning strategies. Bold values indicate the best result for each model and language pair.

Model	**Language Pair**	**Poems FT**	**Poems+Lyrics FT**	**Pre-Trained**
BLEU				
mBART	FR-EN	0.2852	**0.3034**	0.2869
	FR-PT	0.1545	**0.1699**	0.0648
	EN-FR	**0.1546**	0.1253	0.1179
	EN-PT	0.1245	**0.1288**	0.1130
	PT-FR	0.1241	**0.1658**	0.0809
	PT-EN	**0.2528**	0.2059	0.1585
MarianMT	FR-EN	0.2245	0.2197	**0.2741**
	FR-PT	0.1399	0.1007	**0.1611**
	EN-FR	0.1513	0.1240	**0.1746**
	EN-PT	0.1303	0.1057	**0.1404**
	PT-FR	0.1298	0.0355	**0.1791**
	PT-EN	0.1525	0.0898	**0.2344**
OpenNMT (RNN)	FR-EN	0.0324	0.0050	**0.0654**
	FR-PT	0.0242	0.0089	**0.0293**
	EN-FR	0.0212	0.0054	**0.0224**
	EN-PT	**0.0137**	0.0079	0.0131
	PT-FR	0.0357	0.0087	**0.0483**
	PT-EN	0.0429	0.0072	**0.0526**
METEOR				
mBART	FR-EN	0.5470	**0.5779**	0.5590
	FR-PT	0.3737	**0.4120**	0.2385
	EN-FR	**0.3777**	0.3694	0.3597
	EN-PT	**0.3906**	0.3886	0.3769
	PT-FR	0.2856	**0.3693**	0.2511
	PT-EN	**0.5166**	0.4868	0.4259
MarianMT	FR-EN	0.4732	0.5078	**0.5630**
	FR-PT	0.3878	0.3453	**0.4194**
	EN-FR	0.4044	0.3786	**0.4175**
	EN-PT	0.4087	0.3673	**0.4162**
	PT-FR	0.3501	0.2356	**0.3749**
	PT-EN	0.4680	0.3334	**0.5106**
OpenNMT (RNN)	FR-EN	0.1724	0.0509	**0.2712**
	FR-PT	0.1896	0.0996	**0.1995**
	EN-FR	**0.1808**	0.0563	0.1802
	EN-PT	0.1733	0.0907	**0.1872**
	PT-FR	0.2099	0.0673	**0.2130**
	PT-EN	0.2252	0.0567	**0.2886**

Source: Own authorship.

Table 32 – Best automatic evaluation metrics for poetry translations across specialized MT models.

Metric	Language Pair	mBART	MarianMT	Google Translator	OpenNMT (RNN)
BLEU	FR-EN	0.3034	0.2741	**0.3939**	0.0654
	FR-PT	0.1699	0.1611	**0.1960**	0.0293
	EN-FR	0.1546	0.1746	**0.2371**	0.0224
	EN-PT	0.1288	0.1404	**0.1669**	0.0137
	PT-FR	0.1658	0.1791	**0.2124**	0.0483
	PT-EN	0.2528	0.2344	**0.3480**	0.0526
METEOR	FR-EN	0.5779	0.5630	**0.6542**	0.2712
	FR-PT	0.4120	0.4194	**0.4600**	0.1995
	EN-FR	0.3777	0.4175	**0.4902**	0.1808
	EN-PT	0.3906	0.4162	**0.4441**	0.1872
	PT-FR	0.3693	0.3749	**0.4174**	0.2130
	PT-EN	0.5166	0.5106	**0.6191**	0.2886

Source: Own authorship.

Table 33 – Automatic evaluation metrics (BLEU, METEOR, and BERTScore) for poetry translations across LLMs using the same test set as specialized MT models.

Metric	Language Pair	ChatGPT		Maritaca	
		Prompt 1	Prompt 2	Prompt 1	Prompt 2
BLEU	FR-EN	**0.3812**	0.3318	0.3371	0.2911
	FR-PT	0.2421	0.2321	0.2351	**0.3387**
	EN-FR	**0.2149**	0.2026	0.1902	0.2119
	EN-PT	0.1868	0.1760	0.1690	**0.2862**
	PT-FR	0.2531	0.2415	0.2297	**0.2670**
	PT-EN	0.3162	0.2706	0.2987	**0.3321**
METEOR	FR-EN	**0.6448**	0.6125	0.6141	0.5606
	FR-PT	0.4999	0.4953	0.4960	**0.5668**
	EN-FR	**0.4802**	0.4648	0.4609	0.4681
	EN-PT	0.4730	0.4506	0.4502	**0.5248**
	PT-FR	0.4522	0.4385	0.4379	**0.4609**
	PT-EN	**0.6036**	0.5722	0.5896	0.5947
BERTScore	FR-EN	**0.9293**	0.9191	0.9230	0.9076
	FR-PT	0.8106	0.8097	0.8025	**0.8340**
	EN-FR	**0.8105**	0.7969	0.8082	0.8025
	EN-PT	0.7970	0.7908	0.7957	**0.8183**
	PT-FR	**0.7997**	0.7944	0.7876	0.7880
	PT-EN	0.9222	0.9156	0.9209	**0.9236**

Source: Own authorship.