International Journal of Computer Techniques – IJCT Volume 12 Issue 5, October 2025
Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

A Comparative Study of Machine Translation Tools for English-to-Marathi Language Translation

Taiba Imtiyaz Tamboli
Department of Computer Applications
Sinhgad Institute of Business Administration and Research
Pune, India
taibatamboli2410@gmail.com

Simran Mohan Jadhav
Department of Computer Applications
Sinhgad Institute of Business Administration and Research
Pune, India
jadhavsimran111@gmail.com

Assistant Prof. Rubina Sheikh
Department of Computer Applications
Sinhgad Institute of Business Administration and Research
Pune, India
rubina.sk@gmail.com

***Abstract*— The rapid advancement of artificial intelligence has led to the development of highly efficient machine translation systems such as Google Translate, ChatGPT and Gemini. This study presents a comparative evaluation of three machine translation systems — Google Translate, ChatGPT and Gemini — based on both manual and automated assessment methods. Manual evaluation focused on accuracy, fluency, and adequacy, while automated evaluation was conducted using the BLEU (Bilingual Evaluation Understudy) score, calculated through Python programming. A One-Way ANOVA test revealed a statistically significant difference among the BLEU scores ($F(2,147) = 10.26109$, $p < 0.05$), indicating that translation quality varied across systems. Tukey's post-hoc test showed that Gemini performed significantly better than both Google and ChatGPT, while no significant difference was observed between Google and ChatGPT. Error analysis further supported these results, with Gemini showing minimal grammatical errors, Google displaying moderate lexical and idiomatic issues, and ChatGPT exhibiting varied misformation and semantic errors. Overall, Gemini demonstrated the highest translation accuracy and fluency, validating its superior performance among the three systems.**

***Keywords*— Machine Translation, English-Marathi, ChatGPT, Gemini, Google Translate, Error Analysis**

I. INTRODUCTION

In an increasingly globalized world, the demand for accurate and efficient translation systems has risen significantly. Machine Translation (MT) plays a vital role in bridging linguistic barriers, enabling wider access to information, education, and digital communication. With rapid advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP), tools such as Google Translate, ChatGPT and Gemini have become prominent for multilingual translation tasks, each utilizing distinct AI-driven architectures and approaches.

However, translating into regional languages like Marathi—spoken by over 83 million people in India—poses unique challenges due to its Devanagari script, complex ligatures, diacritical marks, and contextual grammar structure. These linguistic features often reduce the performance accuracy of machine translation systems.

This research conducts a comparative analysis of Google Translate, ChatGPT and Gemini for English-to-Marathi translation, combining manual and automated evaluations. Manual assessment focuses on accuracy, fluency, and adequacy, while automated evaluation employs the BLEU (Bilingual Evaluation Understudy) score using Python. Additionally, error analysis categorizes linguistic, idiomatic, and grammatical issues to provide deeper insight into translation performance.

The primary goal of this research is to perform a comprehensive comparative analysis of three major machine translation (MT) systems—Google Translate, ChatGPT and Gemini—specifically for the English-to-Marathi language pair.

The study aims to achieve the following specific objectives:

1. To Quantify Translation Accuracy: To calculate and compare the automated BLEU (Bilingual Evaluation Understudy) scores for the translations generated by Google Translate, ChatGPT, and Gemini, using the NLTK library in Python.

2. To Determine Statistical Significance: To employ a One-Way Analysis of Variance (ANOVA) test to determine if there is a statistically significant difference in the mean BLEU scores among the three MT systems.

3. To Identify Pairwise Differences: To use Tukey's post-hoc test to identify which specific MT system, if any, performs significantly better or worse than the others.

4. To Qualitatively Assess Translation Quality: To conduct a detailed manual evaluation of the translations based on the three key parameters of Accuracy, Fluency, and Adequacy.

5. To Diagnose Systemic Errors: To perform a qualitative error analysis by categorizing specific linguistic and lexical mistakes (including Semantic, Idiomatic, Lexical, and Misformation errors) made by each system to understand their unique systemic weaknesses.

6. To Determine Superior Performance: To conclude which of the three AI-driven MT systems provides the most

reliable, accurate, and fluent translation for the low-resource regional language, Marathi.

II. Literature Review

Machine Translation (MT) has evolved significantly over the past few decades, transitioning from rule-based and statistical models to more sophisticated neural and AI-driven systems. The growing dependency on MT tools such as Google Translate, ChatGPT, and Gemini highlights the need for ongoing evaluation to assess their accuracy, fluency, and contextual understanding across diverse linguistic structures.

Early studies in MT evaluation focused on the limitations of rule-based systems and their inability to handle semantic nuances. Aiken and Balan (2011) conducted an early analysis of Google Translate's accuracy and observed that while it performed acceptably for basic sentence structures, it struggled with idiomatic expressions and complex grammatical forms [1]. This study emphasized the inherent shortcomings of early machine translation systems in capturing linguistic and cultural contexts.

To build upon this, Ghasemi and Hashemian (2016) performed a comparative analysis of Google Translate's English–Persian and Persian–English translations using an error analysis framework. Their findings indicated frequent grammatical, lexical, and semantic inaccuracies, suggesting that Google Translate's neural translation model required refinement for low-resource languages [2]. These results highlighted the persisting gap in translation quality, especially for languages with less digital representation.

A major milestone in MT research was the introduction of the BLEU (Bilingual Evaluation Understudy) metric by Papineni et al. (2002), which provided an automated and standardized method to evaluate MT output against human reference translations [3]. BLEU became a foundational evaluation tool for translation quality, enabling large-scale quantitative comparisons among systems.

Further improvements in translation modeling came with the integration of neural networks. Sennrich, Haddow, and Birch (2016) proposed the use of subword units in neural machine translation, which significantly improved the translation of rare and unseen words [4]. This innovation addressed a major limitation in traditional MT systems and established the foundation for modern neural-based tools such as Google Translate and Gemini.

For regional languages, Goyal and Lehal (2008) developed a Hindi–Marathi machine translation system using a rule-based approach [5]. Their work demonstrated that while rule-based models could provide grammatical consistency, they often failed to produce contextually natural translations. This underlined the need for hybrid and neural approaches that could balance linguistic structure with semantic understanding.

Joshi, Singh, and Saini (2021) carried out a comparative study of multiple MT tools for Indian languages and reported that the neural systems performed better than statistical and rule-based models, especially in terms of fluency and syntactic correctness [6]. However, they also identified a need for localized corpora to improve translation adequacy in Indian contexts.

Evaluation methodologies have also evolved over time. Popović (2015) conducted a comparative analysis of various MT evaluation metrics and concluded that while automatic metrics like BLEU are effective for large datasets, they may not fully capture linguistic nuances or semantic correctness [7]. This reinforced the importance of supplementing automated scores with manual qualitative assessment.

Similarly, Loffler (2021) examined different error typologies in MT and proposed a structured framework for identifying translation errors related to semantics, grammar, and coherence [8]. Such typologies play a crucial role in qualitative error analysis, forming the basis for manual evaluation in the current study.

Recent research has focused on AI-driven translation tools. A 2023 comparative assessment between ChatGPT and Google Translate found that ChatGPT exhibited superior fluency and contextual awareness, particularly in complex sentence structures and idiomatic usage [9]. Another 2024 study comparing Gemini, ChatGPT and Google Translate in translating English idioms into Arabic revealed that Gemini and ChatGPT handled cultural and idiomatic nuances more effectively than Google Translate [10]. These findings indicate a shift toward more human-like translation quality enabled by large language models (LLMs).

Further, Ranathunga et al. (2023) provided a comprehensive overview of the challenges in translating contextually rich, low-resource languages and highlighted the limitations of current NMT systems in capturing domain-specific meaning [11]. Similarly, Haddow et al. (2022) surveyed low-resource MT systems and emphasized the need for adaptive training techniques and multilingual datasets to improve translation performance [12].

Overall, the literature demonstrates a clear progression in machine translation—from early rule-based methods to neural and AI-enhanced systems that integrate contextual understanding. Despite these advancements, persistent issues such as domain adaptation, idiomatic interpretation, and cultural sensitivity continue to affect translation quality. Hence, a comparative evaluation of Google Translate, ChatGPT, and Gemini—using both quantitative BLEU-based metrics and qualitative manual analysis—remains essential to understanding the evolving landscape of MT performance.

III. Methodology And Data Source

This research employed a mixed-methods approach, combining quantitative statistical analysis with qualitative linguistic assessment to provide a robust evaluation of the three machine translation systems.

A. Data and Systems

Target Systems: The study analyzed translations from Google Translate, ChatGPT and Gemini.

Language Pair: The direction of translation was consistently English-to-Marathi.

Sample Size: A common corpus of 50 English source sentences was used, resulting in 50 translated outputs from each of the three systems, totalling N=150 translation segments for analysis.

B. Evaluation Methods

Manual Evaluation (Qualitative)

The translated Marathi outputs were subjected to manual assessment based on the following three linguistic parameters, rated by human evaluators:

1. Accuracy: The degree to which the translated text conveyed the precise meaning of the source text.
2. Fluency: The smoothness and grammatical correctness of the Marathi translation, assessing whether it sounded natural to a native speaker.
3. Adequacy: The extent to which the translation was semantically equivalent and complete, ensuring no loss or distortion of information.

Automated Evaluation (Quantitative)

The quantitative measure used to evaluate the quality of translation was the BLEU (Bilingual Evaluation Understudy) score.

1. Calculation: The BLEU scores were calculated for each translated sentence using the NLTK (Natural Language Toolkit) library in Python, comparing the machine output against human reference translations.
2. Statistical Analysis: The average BLEU scores for all three systems were analyzed using a One-Way Analysis of Variance (ANOVA) test. Goal: To test the null hypothesis (H0): There is no significant difference in mean BLEU scores among the three systems. Parameters: The ANOVA was conducted with 2 degrees of freedom between groups and 147 degrees of freedom within groups (F(2,147)).
3. Post-hoc Testing: Following the significant ANOVA result ($p<0.05$), Tukey's Honestly Significant Difference (HSD) post-hoc test was applied for specific pairwise comparisons to isolate which pairs of systems demonstrated statistically significant differences.

C. Error Analysis

A detailed linguistic error analysis was performed on the 50 translated outputs from each system. Errors were classified into five categories as shown in Table 1. to diagnose specific systemic weaknesses.

Table I. Linguistic Error Categories in Translation

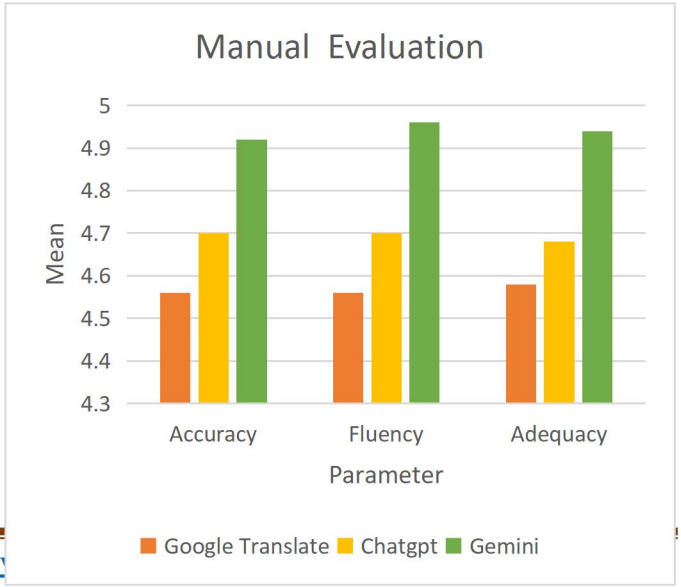
Sr No.	Error Type	Description
1.	Semantic	Errors where the meaning of the source sentence was changed or distorted.
2.	Idiomatic	Errors related to the unnatural translation of cultural phrases, proverbs, or expressions.
3.	Lexical	Errors involving the incorrect choice of individual words or vocabulary.
4.	Misformation	Errors in grammar, morphology, syntax, or word form (e.g., incorrect tense or inflection).
5.	None	Instances where the translation was judged to be accurate, fluent, and adequate.

IV. Results and Discussion

This section presents the findings from both the manual and automated evaluations and discusses their implications with respect to the performance of Google Translate, ChatGPT and Gemini for English-to-Marathi translation.

A. Manual Evaluation Results

The manual evaluation of translation quality was carried out based on three key parameters- Accuracy, Fluency, and Adequacy. The results, as illustrated in Fig.1, indicate that Gemini consistently achieved the highest scores across all three parameters, followed by ChatGPT and Google Translate. Gemini's superior performance suggests more contextually appropriate and grammatically accurate translations. While ChatGPT maintained moderate consistency, Google Translate exhibited slightly lower scores, particularly in fluency and



adequacy. Overall, the manual analysis demonstrates that Gemini provides more natural and coherent translations compared to the other two systems.

Fig. 1. Manual Evaluation Results of Translation Systems

B. Automated Evaluation Results

The automated evaluation, shown in *Fig. 2*, was based on the average BLEU scores calculated using Python. The pie chart indicates that Gemini achieved the highest average BLEU score of 41%, followed by ChatGPT with 32%, and Google Translate with 27%. These results align closely with the manual assessment findings, confirming Gemini's superior translation quality. The BLEU score analysis quantitatively supports that Gemini produces translations that are more similar to human reference translations, validating its higher linguistic accuracy and reliability.

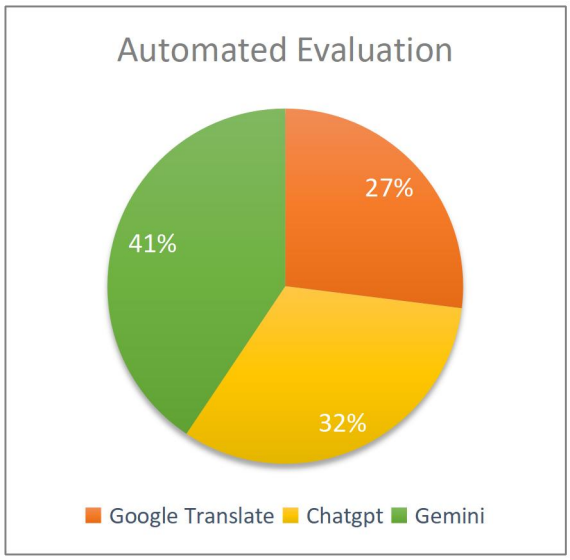


Fig. 2. Result of Automated Evaluation of Translation Systems

C. Qualitative Error Analysis

The error analysis provides crucial insights into the qualitative differences between the three machine translation (MT) systems: Google Translate, ChatGPT and Gemini. This qualitative assessment complements the quantitative BLEU and ANOVA results by categorizing the specific types of linguistic and lexical errors made by each system.

Overall Error Distribution

The bar chart fig. 3, visually summarizes the distribution of errors across five primary categories: Semantic, Idiomatic, Lexical, Misformation and None (no error).

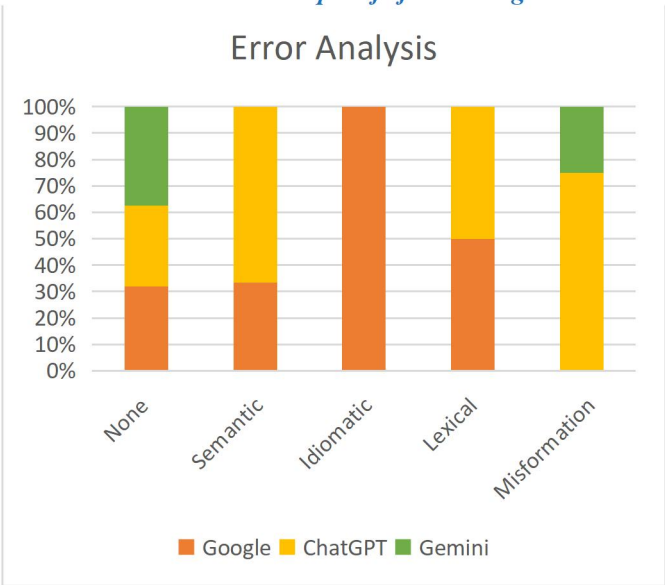


Fig. 3. Categorical Error Analysis of MT Systems

Lowest Total Error: The system with the lowest total error rate (i.e., the highest percentage in the 'None' category) is Gemini. Gemini's translations were rated as having "None" (no error) approximately 48% of the time, compared to Google's approximately 41% and ChatGPT's approximately 39%. This aligns with the quantitative finding that Gemini's overall performance (as measured by BLEU and Tukey's test) was significantly superior.

Gemini's Superiority: The qualitative analysis strongly supports the quantitative findings, as Gemini produced the fewest errors overall (only 2 errors out of 50 translations). Notably, Gemini produced zero Semantic, Idiomatic, and Lexical errors.

Systematic Weaknesses:

- Google Translate's primary weakness was in Idiomatic errors (6 instances). This suggests a struggle to produce translations that sound natural or culturally appropriate for the target language.
- ChatGPT demonstrated a primary vulnerability in Misformation errors (6 instances). These errors involve incorrect grammatical structure or word forms, indicating a lack of consistent morphological control.
- Semantic Errors were exclusively found in the output of Google (2 instances) and ChatGPT (4 instances), and were entirely avoided by Gemini. Semantic errors are the most critical, as they alter the original meaning, suggesting a major strength for Gemini in preserving core content.

V. Statistical Analysis

To evaluate the translation performance of Google Translate, ChatGPT and Gemini, both manual and automated assessments were conducted. Manual evaluation considered accuracy, fluency, and adequacy, while automated evaluation used the BLEU (Bilingual Evaluation Understudy) score, calculated with Python. A one-way ANOVA was applied to determine whether the mean BLEU scores differed across the three systems. Formally, the null hypothesis is as follows:

ANOVA H_0(null): There is no significant difference in mean BLEU score accuracy among the systems.

If the null hypothesis is not rejected, the results indicate that, on average, translation accuracy does not differ significantly between the systems. If the null hypothesis is rejected, it suggests that at least one system performs differently, leading to variation in translation quality.

As shown in Table 2, the ANOVA test produced an F-value of 10.26109, which is statistically significant at $p < 0.05$. This finding indicates that there are significant differences in the mean BLEU scores among the three translation systems. Thus, the null hypothesis was rejected, confirming that the translation performance varied notably across the models.

Table 2. Summary of One-Way ANOVA for Translation Quality Scores.

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Between Group	2.04358	2	1.02179306	10.26109
Within Group	14.6381	147	0.09957943	–
Total	–	149	–	–

To identify which systems differed significantly, a Tukey's HSD post-hoc test was conducted, as presented in Table 3. The results revealed that the difference between Google Translate and ChatGPT was not statistically significant, indicating similar translation performance between the two systems. However, significant differences were observed in the pairs Google Translate vs. Gemini and ChatGPT vs. Gemini, suggesting that Gemini's translation accuracy differs significantly from both Google Translate and ChatGPT.

Table 3. Tukey's HSD Post-hoc Test Results for Pairwise Comparisons.

Pairwise	Difference	Significant
Google vs ChatGPT	0.112635207	No
Google vs Gemini	0.283897406	Yes
ChatGPT vs Gemini	0.171262199	Yes

VI. Limitations & Future Scope

A. Limitations of the Study

The current comparative analysis, while comprehensive, was subject to certain limitations that restrict the generalizability of the findings:

- Limited Corpus Size and Domain: The study was conducted using a sample size of only 50 English source sentences across a general domain. This sample may not fully capture the complexity, variability, or stylistic demands of real-world translation tasks, nor can it ensure complete coverage of all possible linguistic features of the Marathi language.
- BLEU Score Dependency: While standard, the BLEU score is primarily a precision-focused, n-gram overlap metric. It does not perfectly correlate with human judgment, nor does it penalize semantic errors severely, which may slightly skew the quantitative results toward fluency over deep meaning.
- Black-Box Nature of LLMs: The exact current training data, model architecture (e.g., GPT-3.5 vs. GPT-4 for ChatGPT, or the specific Gemini version), and internal workings of the commercial systems (Google Translate, ChatGPT, Gemini) are proprietary. This makes it impossible to definitively pinpoint the exact cause of the observed error patterns.
- Language Directionality: The research focused exclusively on English-to-Marathi translation. The performance hierarchy of the three systems may be entirely different for the reverse direction (Marathi-to-English) or for other regional Indian language pairs.

B. Future Scope of the Study

This study showed that Gemini is the best tool for English-to-Marathi translation, but all systems can be improved. Future research should focus on the following areas:

- Fixing Specific Errors: Developers should focus on helping Google Translate handle Marathi idioms and common phrases better. Work should be done to reduce the grammatical mistakes (Misformation errors) often made by ChatGPT.
- Better Testing Methods: Instead of just using the BLEU score, future studies can use other quality tests like METEOR or COMET to achieve higher accuracy. The tools should be tested on different types of Marathi text (e.g., medical, legal, or

technical documents) to see which one works best for specific jobs.

- Real-World Use: Researchers should study how much time and effort a human translator needs to fix the output of each machine (Google, ChatGPT, Gemini). This will tell us which tool is the most practical to use in a professional setting. More high-quality Marathi translation data needs to be created to help train future, more accurate AI models.

VII. Conclusion

This paper presented a comprehensive comparative analysis of Google Translate, ChatGPT, and Gemini for English-to-Marathi machine translation using both manual and automated evaluation methods. Manual assessment based on accuracy, fluency, and adequacy revealed that Gemini consistently produced more natural and contextually appropriate translations, followed by ChatGPT and Google Translate.

Automated evaluation using the BLEU score, computed through Python, supported these findings. The one-way ANOVA test yielded an F-value of 10.26109, indicating statistically significant differences among the mean BLEU scores of the three systems. Subsequent Tukey's post-hoc pairwise analysis showed that while the difference between Google and ChatGPT was not significant, both Google vs. Gemini and ChatGPT vs. Gemini comparisons were statistically significant. This confirms Gemini's superior translation quality in terms of automated evaluation metrics.

The error analysis further revealed that Gemini produced the fewest linguistic and semantic errors, while ChatGPT exhibited frequent lexical and grammatical misformations. Google Translate demonstrated balanced performance but occasionally struggled with idiomatic and context-dependent phrases.

Overall, the findings highlight that Gemini outperforms Google Translate and ChatGPT in both manual and automated assessments for English-to-Marathi translation. The study emphasizes the growing reliability of advanced AI-driven systems like Gemini and ChatGPT in handling low-resource regional languages, while also identifying areas for future improvement in context understanding and idiomatic accuracy.

VIII. References

[1] M. Aiken and Sh. Balan, "An analysis of Google Translate accuracy," Translation Journal, vol. 16, no. 2, 2011.

[2] H. Ghasemi and M. Hashemian, "A comparative study of Google Translate translations: An error analysis of English-to-Persian and Persian-to-English translations," English Language Teaching, vol. 9, no. 3, pp. 13–17, 2016.

[3] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. 40th Annual Meeting on Association for Computational Linguistics (ACL), 2002, pp. 311–318.

[4] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016, pp. 1715–1725.

[5] P. Goyal and G. S. Lehal, "Hindi–Marathi machine translation system using rule-based approach," Journal of Computing and Information Technology, vol. 16, no. 3, pp. 225–232, 2008.

[6] M. Joshi, S. Singh, and S. Saini, "Comparative study of machine translation tools for Indian languages," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 12, no. 3, pp. 45–52, 2021.

[7] M. Popović, "A comparative analysis of evaluation metrics for machine translation," Language Resources and Evaluation, vol. 49, no. 3, pp. 679–705, 2015.

[8] E. Loffler, "Error typologies for machine translation quality assessment," Linguistica Antverpiensia, vol. 20, pp. 230–245, 2021.

[9] P. Rao, L. M. McGee, and C. A. Seideman, "A comparative assessment of ChatGPT vs. Google Translate for the translation of patient instructions," J. Med. Artif. Intell., vol. 6, no. 11, 2023.

[10] M. M. Obeidat, A. S. Haider, S. Abu Tair, and Y. Sahari, "Analyzing the performance of Gemini, ChatGPT, and Google Translate in rendering English idioms into Arabic," FWU J. Soc. Sci., vol. 18, no. 4, pp. 1–18, 2024.

[11] R. Ranathunga, R. Sennrich, et al., "Overview and challenges of machine translation for contextually-rich, low-resource languages," PMC Survey, 2023.

[12] B. Haddow, R. Bawden, A. V. Miceli Barone, J. Helcl, and A. Birch, "Survey of low-resource machine translation," ACL Anthology, 2022.