LINGTERSA
(Linguistik, Terjemahan, Sastra)
Journal homepage: https://talenta.usu.ac.id/lingtersa

Translating the Untranslatable: DeepL and ChatGPT on Academic Idioms

Feby Dora Nurcahyani[*1] , Dimas Adika[2] , Widyasari[3]
[1]*Indonesia Open University, Bogor, 16164, Indonesia*
[2]*Sebelas Maret University, Surakarta, 57126, Indonesia*
[3]*Indonesia Open University, Tangerang, 15437, Indonesia*
[*]Corresponding Author: 042040879@ecampus.ut.ac.id

ARTICLE INFO
E-ISSN: 2964-1713
P-ISSN: 2775-5622

ABSTRACT

This research explores the efficacy of two prominent machine translation platforms, DeepL and ChatGPT, in translating academic idioms from English to Indonesian. Academic idioms, situated between discipline-specific jargon and universally understood expressions, pose a challenge for existing translation systems, particularly those rooted in Neural Machine Translation (NMT). The study employs a qualitative descriptive methodology, focusing on translation precision and naturalness, with bilingual experts evaluating translations through a questionnaire, focusing on translation precision and naturalness. The comprehensive analysis involved 50 participants who assessed translations on a scale of accuracy and fluency using Fiederer and O'Brian's (2009) rating scale. The results indicate that both platforms exhibit strengths and weaknesses in terms of accuracy and fluency. While DeepL demonstrates trust in its translation proficiency, ChatGPT receives a more favorable response, especially regarding fluency. Participants preferred ChatGPT for fluency in handling academic expressions, indicating its adaptability. The study also revealed a general agreement among participants regarding the difficulties both platforms encounter in accurately translating academic idioms, emphasizing continuous requirements for improved machine translation. These insights enhance understanding of machine translation's strengths and limitations in academic setting, with implications for future technology development.
Keyword: Academic Idioms, ChatGPT, DeepL , Translation Accuracy

ABSTRAK

Penelitian ini mengeksplorasi efektivitas dua platform penerjemahan mesin, DeepL dan ChatGPT, dalam menerjemahkan idiom akademis dari Bahasa Inggris ke Bahasa Indonesia. Idiom-idiom akademis, yang berada di antara disiplin disiplin ilmu dan ekspresi yang bisa dipahami secara universal, menjadi tantangan tersendiri bagi sistem penerjemahan yang sudah ada, terutama yang berbasis *Neural Machine Translation* (NMT). Studi ini menggunakan metodologi kualitatif, yang fokus pada keakuratan dan kealamiahan terjemahan, para responden yang menguasai dwibahasa mengevaluasi terjemahan melalui kuesioner. Hasil penelitian menunjukkan bahwa kedua platform menunjukkan kelebihan dan kelemahan pada keakuratan dan kealamiahan DeepL dan ChatGPT. DeepL menunjukkan keakuratan terjemahannya, namun ChatGPT mendapatkan respon yang yang lebih baik, terutama terkait keakuratan dan kealamiahan. Studi ini menggarisbawahi meningkatnya ketergantungan pada penerjemahan mesin dalam konteks akademis dan menyoroti perlunya perhatian lebih pengembangan penerjemahan idiom yang bernuansa akademis secara akurat.
Keyword: Academic Idioms, ChatGPT, DeepL , Translation Accuracy

http://doi.org/10.26594/register.v6i1.idarticle

1. Introduction

For accurate translation, language's complexity necessitates an intricate comprehension of semantics, context, and culture. Several machine translation systems have attempted to reconcile the language gap over the years. Early models, constructed on the basis of principles and statistics, were only marginally successful. Rule-based machine translation (RBMT) and statistical machine translation (SMT) were the leaders prior to Neural Machine Translation (NMT). Brown et al. (1993) found that while RBMT relied on linguistic norms and dictionaries, SMT relied on statistical analyses of bilingual text corpora. Both these systems, despite being pioneering, had their limitations, particularly in grasping and conveying nuanced contextual meanings.

Transitioning from rules and statistics, Neural Machine Translation NMT, grounded in deep learning, utilizes neural networks to decipher intricate semantic and contextual relationships between words in vast datasets. This innovation led to translations of superior quality, offering a deeper understanding of linguistic nuances. The transformer model, introduced by Vaswani et al. (2017), further advanced this field, giving birth to numerous models designed to better interpret human language.

Google Translate embarked on a transformative journey towards NMT in 2016, as documented by Wu et al (2016). Their NMT model demonstrated a significant reduction in translation errors, frequently generating outputs that are more natural and fluent. Google Translate has emerged as the preferred choice for novice individuals who seek translation services. Launched in 2017, DeepL was a machine translation service purposefully designed from the outset to harness the capabilities of Neural Machine Translation (NMT). With its foundation rooted in advanced neural networks and access to a comprehensive corpus of multilingual data, DeepL quickly garnered widespread attention for its superior translation abilities. Significantly, it demonstrated a distinct aptitude in accurately capturing language subtitles and idiomatic expressions, sometimes surpassing other well-established translation tools, such as Google Translate. The exceptional ability of DeepL's algorithms to produce translations that are very fluent and resonate intuitively with native speakers was notably showcased in a 2017 article published by the esteemed technology magazine, Wired. The platform was discovered to surpass existing online translation services in terms of both quality and accuracy.

Moreover, Vaswani et al. (2017) conducted a study which revealed that DeepL's translations exhibited superior performance compared to human translators in specific scenarios, thereby highlighting its exceptional potential. The emotion expressed above was similarly conveyed in a comparative analysis conducted by Schwenk et al. (2019). In this study, DeepL demonstrated competitive or, in certain cases, better performance compared to other neural machine translation (NMT) models in different language contexts.

The concentrated design approach adopted by DeepL, which is only dedicated to translation, has allowed for the refinement and enhancement of both the precision and fluency of its translated outputs. The work presented in the Proceedings of the Association for Computational Linguistics (2019) further emphasized the precision in design and intentionality. The specific design of DeepL was highlighted for its notable performance benefits compared to more general models, particularly in translations that need a comprehensive grasp of context.

Despite its intended purpose not being focused on translation, chatGPT shows the adaptability and effectiveness of Neural Machine Translation (NMT). The platform demonstrates exceptional proficiency in comprehending and generating content that has human-like structures in several languages as evidenced by the publication "Language Models are Few-Shot Learners" (2020) by OpenAI. ChatGPT which is built upon the foundational GPT-3 model demonstrates a high level of competence in many language tasks, such as translation. This proficiency is observed even when the model is provided with a little amount of task-specific training data, highlighting its versatility and capacity to produce translations of superior quality.

The validation of the inherent strength in retaining conversational settings, particularly in cross-language encounters, was conducted by Radford et al. (2019). Furthermore, a multitude of comparative assessments have depicted a positive outlook on the capabilities of ChatGPT. Significantly, previous research conducted by Zulfiqar et al. (2018) and Araújo & Aguiar (2023) has suggested that ChatGPT has promising capabilities in the domain of specialized translation scenarios. In their research conducted in 2023, Sanz-Valdivieso and López-Arroyo (2023) focused their attention on specialized terminology inside niche translation domains. The results of their study revealed an increasing inclination towards utilizing ChatGPT as opposed to other platforms, such as Google Translate. In a scholarly investigation on the translation of scientific texts from English to Portuguese, Araújo and Aguiar (2023) discovered that the translations produced by ChatGPT consistently received superior ratings in terms of fluency, accuracy, appropriateness, and overall evaluation when compared to assessments conducted by human evaluators.

Additional comparisons conducted inside specialist translation contexts yielded intriguing results. In a recent study conducted by Lucia Sanz-Valdivieso and López-Arroyo (2023), it was determined that ChatGPT exhibits a higher level of performance compared to Google Translate. Notably, the researchers saw a reduction in terminological mistakes while utilizing ChatGPT. Furthermore, Calvo-Ferrer (2023) highlighted the challenge faced by viewers in distinguishing between subtitles generated by ChatGPT and those created by human translators in English- Spanish translations. This observation highlights the significant progress in ChatGPT's language proficiency over time. In addition to conventional settings, Zhao et al. (2023) noted that when using neural machine translation (NMT) systems to literary material, platforms such as ChatGPT exhibit a more extensive lexicon and enhanced performance metrics compared to NMT systems designed for general purposes. Additionally, a particular mistake categorization specifically designed for literary translation was established by them.

Then, a comprehensive review undertaken by Jiao, Wang, Huang, Wang, and Tu (2023), the effectiveness of ChatGPT-4 as a translation engine was thoroughly examined, further solidifying its standing in the field of translation. The aforementioned feeling was reiterated by Castilho, Sheila, Mallon, Clodagh, Meister, Raheel, and Yue, Shenghua (2023), whose research revealed that the GPT model exhibits superior performance compared to other neural machine translation (NMT) systems, however with some limited cases where this superiorit is not seen.

The recent advancements in the field of machine translation have seen notable progress, higher performance exhibited by DeepL and OpenAI's ChatGPT in comparison to systems such as Google Translate, but they have trouble translating academic words that are strongly rooted in certain fields (Li & Zou, 2019). Academic idioms are frequently situated between discipline-specific jargon and universally understood expressions. Dankers, Lucas, and Titov (2022) indicate that Neural Machine Translation (NMT) models, especially those based on the Transformer architecture, struggle with these idiomatic expressions and frequently treat them as singular entities. ChatGPT, despite being predominantly a text generator (Brown et al., 2020), exhibits multilingual capabilities and has undertaken translation duties. Its training on diverse datasets has provided it with a comprehensive comprehension of languages, but translating academic idioms, with their nuanced meaning, remains a challenging task.

Due to the high expense of human translation services and people from different backgrounds demand information for various reasons (Anggawijaya & Adika, 2023), the academic community is increasingly relying on automatic machine translation as alternative. The increasing popularity of tools like as DeepL and ChatGPT can be attributed to their rapid processing capabilities and ongoing improvements. However, the presence of academic idioms, which include intricate connotations, continues to pose a substantial challenge for these sophisticated platforms. One misconception of idioms is the idea that it is actually not appropriate to use idioms in academic English since English is supposed to be formal academic English. The idioms in academic English, such as information about what an idiom is and the reasons why academic idioms should be studied and a list of academic idioms for spoken and written English from a latest study of idioms (Miller, 2019) providing context for the list's creation and then the list itself. Studies conducted by Simpson and Mendis (2003) and Miller (2019) discovered the use of idioms in academic environments, which runs counter to the common perception that idioms are informal forms of speech. Understanding idioms can help students better integrate into academic discourse, since Miller's study showed that their prevalence in academic writings was 0.1%.

This study concentrates on translating academic idioms from English to Indonesian in order to evaluate the efficacy popular machine translation platforms, DeepL and chatGPT. With increasing academic and cultural exchanges between English and Indonesian speakers, it is essential to evaluate how these tools manage idiomatic translations, particularly in light of Indonesian's unique linguistic challenges. The research highlights the increasing reliance on machine translation in academic contexts and the need for accurate translation of academic idioms. Misinterpretation of idiomatic expressions can give rise to misconceptions, hence posing a risk of disseminating inaccurate information within the realm of intellectual discourse. The objective of this study is to address this research gap by providing valuable insights on the strengths and weaknesses of contemporary translation technologies. The idioms employed in this research are derived from a 2019 study by Julia Miller. Miller utilized two primary corpora of academic English: specifically, for the spoken components, the British Academic Spoken English (BASE) was used, and for the written texts, the Oxford Corpus of Academic English (OCAE) was employed. Only idioms with a frequency greater than 1.2 occurrences per million words (pmw) in the BASE corpus were considered. The range of idiom usage is indicated by the number of texts and faculties in which each idiom appears. The extent of idiom utilization is demonstrated by the frequency of its occurrence in various texts, as well as its presence across many faculties or academic groupings. The idioms included in the list exhibit a cross-disciplinary nature,

rendering them appropriate for examination by students across many academic disciplines studying English. Miller's study utilized four such faculties, namely Social Sciences (234 pmw), Arts and Humanities (191 pmw), Life and Medical Sciences (183 pmw), and Physical Sciences (76 pmw), with Physical Sciences having the least frequent usage. In total, there are 170 spoken idioms, and 38 idiomatic expressions tailored for written academic English are available. The OCAE is a reputable academic corpus, and the idioms included in the list are likely to be well-researched and accurately identified. Academic idioms are used in scholarly writing and research, making them a valuable focus for research**.** They are useful as representative set of idioms within academic contexts.

2. Method

The purpose of this qualitative descriptive, focused on the translation's precision, and naturalness. The research collects information by giving questionnaire of a direct translation of each idioms from each platform and ask the participant to evaluate the accuracy and the fluency. The reviewers are bilingual experts (English and Indonesian) with experience in academic contexts to evaluate translations. The analysis of accuracy is answering does the translation convey the idiomatic meaning or is it overly literal? It focuses on whether the translation preserves the idiomatic substance or tends toward a literal interpretation. Meanwhile, fluency is trying to answer the question does the translation sound natural in Indonesian, especially in an academic writing? The criteria assess whether the translation sounds natural in an academic Indonesian setting.

The translation was evaluated in terms of accuracy and fluency employing Fiederer and O'Brian's (2009) rating scale. Their scale divides accuracy into four distinct categories: highly accurate, accurate, less accurate, and inaccurate. Translation naturalness is the ease of understanding the translation (Fiederer and O'Brian, 2009). The translation is considered to have a high level of naturalness when the target audience can completely comprehend it. Larson (1998 p.529) emphasized that translation naturalness means that the text is easy to understand, which is demonstrated by the target language's appropriate language style. The categories of fluency are highly natural, natural, less natural, and unnatural. It is important to highlight that the study utilized the free versions of these platforms, the ChatGPT 3.5 version and the standard DeepL service. This decision is in line with the pragmatic factor that not all academics may be willing to invest in subscription plans, emphasizing the accessibility and common usage of free versions.

3. Results and Discussions

In the beginning, a pilot study was undertaken to assess the validity of the questionnaire and grading methodology by testing a restricted set of idioms. The initial phase can aid in identifying possible challenges related to clarity for reviewers and allow for the incorporation of necessary adjustments before initiating the core data collection procedure.

We did blind evaluations, in order that reviewers should not be aware of which translation platform (DeepL or ChatGPT) provided each translation. In order to mitigate any bias in the assessment of translations generated by ChatGPT and DeepL, research questioner are presented one by one. By sequentially presenting the translations, reviewers are afforded the opportunity to concentrate their attention on each individual platform. Presenting them side by side may inadvertently introduce bias because the reviewers could compare them directly, which could influence their assessments. By eliminating direct comparisons, we will obtain more objective and unbiased feedback on the quality of each translation. The duration of the evaluation was estimated to be approximately 30 minutes for each section, a pilot run took 10 -15 minutes.

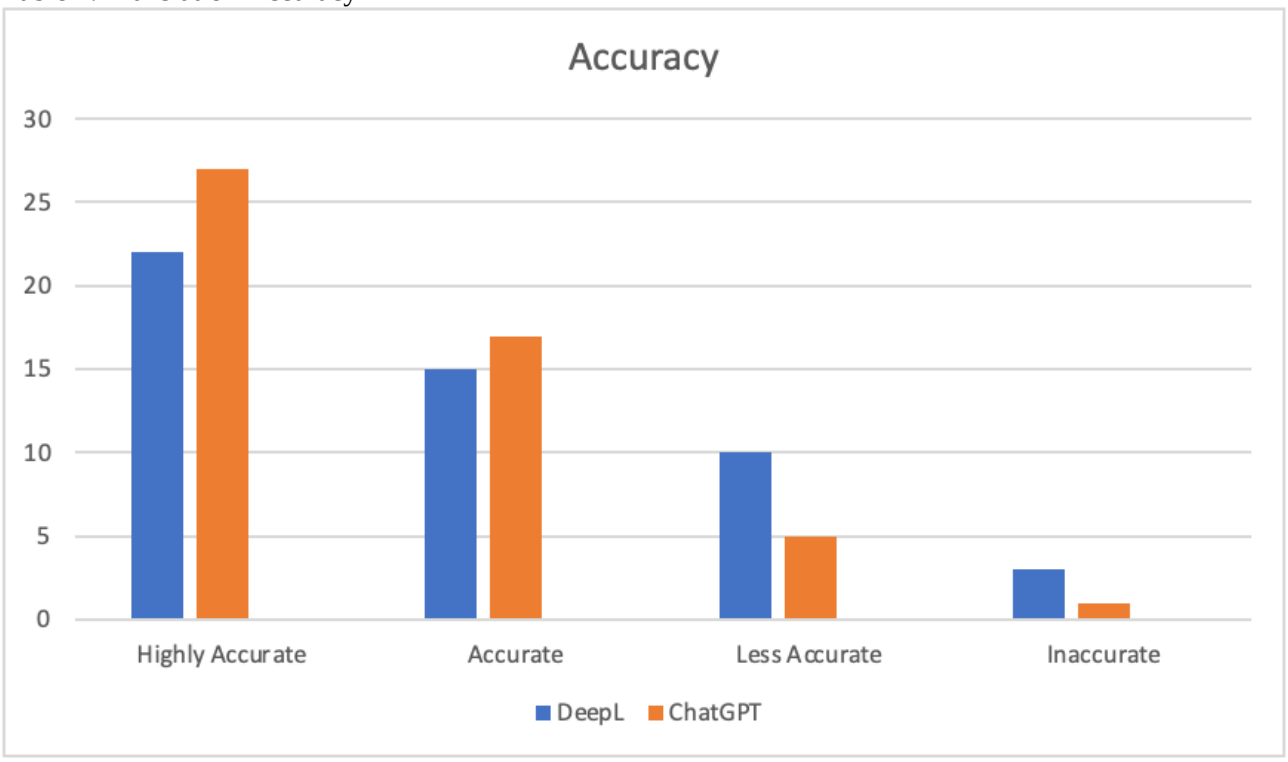
Reviewers' Profile

A total number of 50 evaluators agreed to take part in this research. All evaluators fulfilled the following criteria:

- They were held post-graduate level, some of them are on PhD research
- They had a high level of competence in English, and some of them had studied abroad.

Translation Accuracy

Table 1. Translation Accuracy



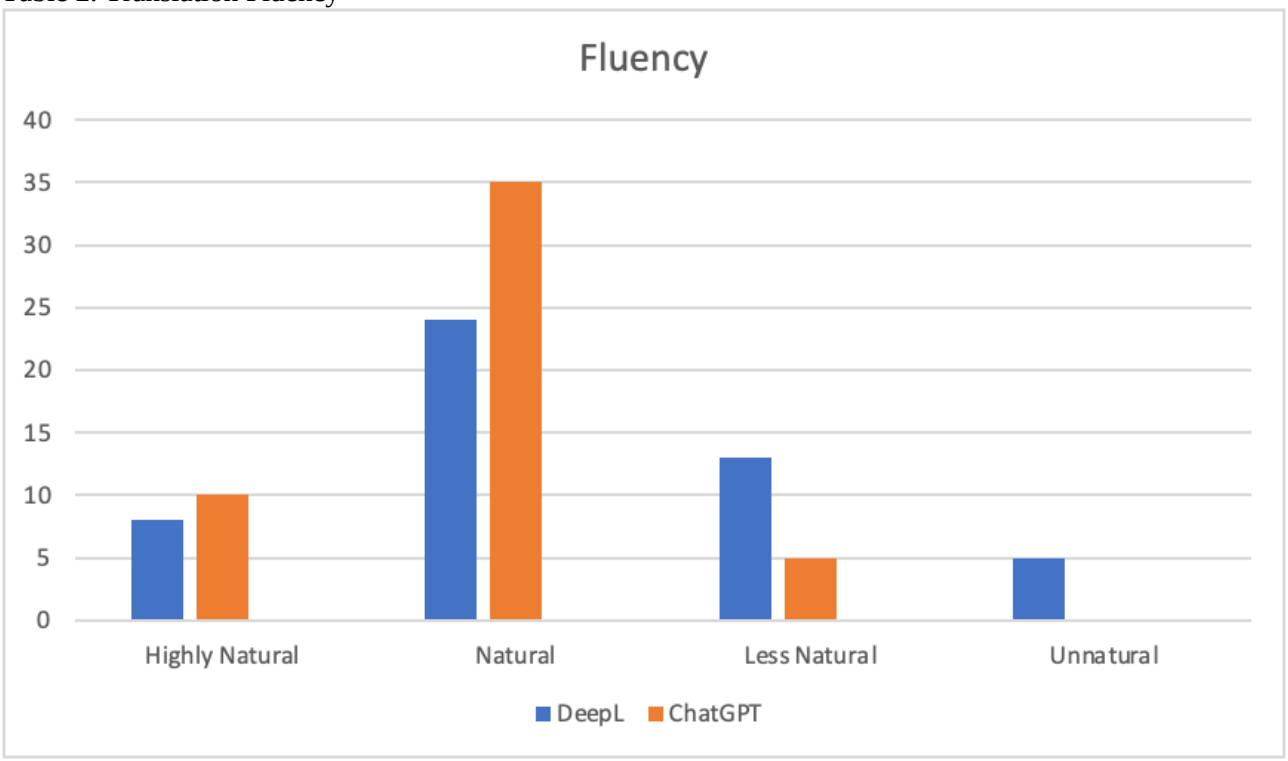
As seen in table 1, the data summaries provide a full comparison examination of translation accuracy between DeepL and ChatGPT, encompassing the four stated criteria. The research results, obtained from a sample size of 50 participants, provide valuable insights into the perspectives of the two translation systems by the participants

Within the framework of DeepL, the data indicates that a total of 22 participants exhibited a considerable degree of trust in the precision of the translations, hence signifying a noteworthy inclination towards positive emotion. In addition, it was observed that 15 participants saw the translations as accurate, whereas 10 participants perceived them to be less precise. A minimum of three participants considered the translations given by DeepL to be categorically incorrect. This comprehensive analysis underscores the diversity of viewpoints concerning the correctness of DeepL, wherein a significant majority of participants demonstrate trust in its translation proficiency.

When considering ChatGPT, the observations suggest a more positive reception. In particular, a total of 27 participants expressed a perception of the idiomatic translation provided by ChatGPT as being highly accurate. This observation indicates a greater degree of trust in the precision of ChatGPT's translations compared to those generated by DeepL. In addition, a total of 17 participants expressed that the translations were typically precise, suggesting a noteworthy favorable pattern. A subset consisting of five participants expressed that the translations exhibited a lesser degree of precision, whilst a sole participant regarded them as erroneous. The collective data indicate a favorable attitude towards ChatGPT in relation to its correctness, as a majority of participants expressed trust in its ability to translate proficiently.

Translation Fluency

Table 2. Translation Fluency



Expanding on the research findings, the nuanced evaluation of translation fluency between DeepL and ChatGPT, derived from data collected from a diverse sample of 50 participants, provides deeper insights into user perceptions and preferences. The criteria evaluate the degree to which the translation exhibits naturalness within the context of an academic environment in Indonesian.

Within the realm of DeepL, the data unveils a spectrum of opinions. While 8 participants exhibited a highly favorable view, perceiving DeepL's translations as "highly natural," a more substantial cohort of 24 respondents still acknowledged a natural quality in the platform's output. This duality in responses highlights the subjective nature of fluency assessment, suggesting that participants might have varying criteria for what constitutes a highly natural translation. Conversely, 13 participants expressed a perception of reduced naturalness, indicating potential areas for improvement, and only a marginal 5 participants deemed the translations as outright unnatural. This distribution of responses underscores the importance of understanding the diverse ways in which users interpret and assess translation fluency.

Turning attention to ChatGPT, a clear trend of positivity emerges. A noteworthy 10 participants perceived the translation of idioms as "highly natural," reflecting a substantial degree of confidence in ChatGPT's fluency. The majority, consisting of 35 respondents, found the translations to be simply "natural," indicating a widespread positive sentiment toward ChatGPT's ability to produce linguistically fluid translations. Furthermore, the relatively low count of 5 participants noting less natural translations suggests a general consensus on the platform's effectiveness. Impressively, none of the respondents considered the translations as outright "unnatural," emphasizing the overall favorable view of ChatGPT's fluency within the sampled group.

Research Insights into ChatGPT, DeepL, and Machine Translation Systems

The research by Yuxin et al. (2023) aligns with our earlier findings, providing a comparative analysis of ChatGPT and widely used machine translation systems. It highlights ChatGPT's superior accuracy, fluency, and logical coherence in translating diverse and complex content, especially in complex scenarios. Additionally, Li et al. (2023) revealed ChatGPT's exceptional proficiency in idiomatic translation.

Examining our research within the existing literature on machine translation and academic language is crucial. We contribute by shedding light on the nuanced challenges and successes observed in translating academic idioms, specifically employing the free versions of ChatGPT 3.5 and standard DeepL service.

Our results align with previous studies, showcasing the remarkable proficiency of machine translation platforms in handling academic language. The high accuracy demonstrated by both DeepL and ChatGPT resonates with findings from Yuxin et al. (2023) and Li et al. (2023), reinforcing the notion that these platforms excel in providing precise translations, even in complex scenarios.

Furthermore, our research delves into the fluency aspect, revealing the subjective nature of user assessments. This resonates with observations by Jiao, Wang, Huang, Wang, and Tu (2023), emphasizing the importance of considering user perceptions in evaluating machine translation systems. Varied opinions on fluency suggest areas for improvement, aligning with the continuous refinement highlighted by Castilho, Sheila, Mallon, Clodagh, Meister, Raheel, and Yue, Shenghua (2023).

The acknowledgment of the practicality and accessibility of free versions in our study echoes sentiments expressed by Sanz-Valdivieso and López-Arroyo (2023) and Araújo and Aguiar (2023), underscoring the increasing preference for platforms like ChatGPT, especially in specialized translation scenarios.

However, our study brings attention to the persistent challenge of translating academic idioms, a theme not extensively explored in prior literature. Dankers, Lucas, and Titov (2022) and Miller (2019) indicated the struggles of NMT models with idiomatic expressions, aligning with our findings. This underscores the need for ongoing advancements in machine translation to address these linguistic intricacies.

While the increasing reliance on machine translation in academic contexts is emphasized, driven by factors such as rapid processing capabilities and ongoing improvements in free platforms, our findings underscore the continual challenge of translating academic idioms. These idioms, rich in nuanced meanings and context, pose a continual obstacle for even advanced machine translation systems.

As machine translation continues to play a pivotal role in facilitating cross-cultural and academic exchanges, there is an evident need for further advancements. The study emphasized the importance of accurate translation in academic discourse, where misinterpretations of idiomatic expressions could lead to the dissemination of inaccurate information.

In conclusion, our research aligns with existing literature, affirming the proficiency of free versions of machine translation platforms in academic language contexts. However, it introduces a novel focus on the translation of academic idioms, highlighting both successes and challenges. The synthesis of our findings with previous studies enriches the broader understanding of the evolving landscape of machine translation in academic settings, emphasizing the importance of addressing nuanced language aspects for continuous improvement.

Conclusions

DeepL impressively demonstrates a high level of trust in its translation proficiency, particularly in terms of accuracy. However, the subjective nature of fluency assessment indicates varying opinions among users, suggesting potential areas for improvement. On the other hand, ChatGPT receives a more favorable response, with a majority of participants expressing high accuracy and fluency in its translations of academic idioms. The platform's proficiency in handling linguistic nuances and producing natural-sounding translations is notable, reflecting its adaptability and effectiveness in specialized translation scenarios.

Looking ahead, future research efforts should delve deeper into specific areas of academic translation that could benefit from improvement. For instance, investigating the challenges posed by certain types of idioms or academic disciplines could guide targeted enhancements. Additionally, synthesizing our findings with existing literature on machine translation and academic idioms can deliver a more comprehensive concept of the evolving landscape.

In conclusion, the findings of this study not only serve as a foundational understanding of the capabilities of machine translation platforms in handling academic idioms but also pave the way for further investigations. The continuous refinement and advancement of machine translation technologies remain integral to bridging language gaps and promoting effective communication in academic settings.

References

Academic Idioms. (n.d.). Www.eapfoundation.com. Retrieved October 29, 2023, from https://www.eapfoundation.com/vocab/academic/other/idioms/index.php?type=written#idiomslist

Anggawijaya, M. H., & Adika, D. (2023). Enhancing Target Text Comprehension for Lay Audience through Paraphrasing. Jurnal Humaya: Jurnal Hukum, Humaniora, Masyarakat, Dan Budaya, 3(1), 1–14. https://doi.org/10.33830/humaya.v3i1.4282

Araújo, S., & Aguiar, M. (2023). Comparing ChatGPT's and Human Evaluation of Scientific Texts' Translations from English to Portuguese. . . *ResearchGate*. https://www.researchgate.net/publication/374061693_Comparing_ChatGPT's_and_Human_Evaluation_of_Scientific_Texts'_Translations_from_English_to_Portuguese_Using_Popular_Automated_Translators_Notebook_for_the_SimpleText_Lab_at_CLEF_2023

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, *19*(2), 263–311. https://aclanthology.org/J93-2003/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33.

Castilho, S., Mallon, C., Meister, R., & Yue, S. (2023, June 1). *Do online machine translation systems care for context? What about a GPT model?* Doras.dcu.ie. https://doras.dcu.ie/28297/

Calvo-Ferrer, J. R. (2023). Can you tell the difference? A study of human vs machine-translated subtitles. *Perspectives*, 1–18. https://doi.org/10.1080/0907676x.2023.2268149

Cheng Yuxin, Wang Ruochen, Chen Jiawei, Chao Yijun, Aliye Maimaitili, & Zhang Haoruo. (2023). Context-Based AI Translation From a Globalization Perspective: A Case Study of ChatGPT. *Sino-US English Teaching*, *20*(9). https://doi.org/10.17265/1539-8072/2023.09.005

Dankers, V., Lucas, C. G., & Titov, I. (2022). Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. *ArXiv:2205.15301 [Cs]*. https://arxiv.org/abs/2205.15301

Fiederer, R. & O'brien, S. (2009). Quality and Machine Translation: A realistic objective?. The Journal of Specialised Translation 11, 52-74

Hariri, W. (2023). Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing. *ArXiv:2304.02017 [Cs]*. https://arxiv.org/abs/2304.02017

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023, February 17). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation*. ArXiv.org. https://doi.org/10.48550/arXiv.2302.09210

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2301.08745

Sanz-Valdivieso, L., & López-Arroyo, B. (2023). Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation?. *HiT-IT 2023*, 97-107.

Katz, M. (n.d.). *Welcome to the Era of the AI Coworker | Backchannel*. Wired. Retrieved October 30, 2023, from https://www.wired.com/story/welcome-to-the-era-of-the-ai-coworker/

Larson, M. L. (1998). *Meaning-based Translation*.

Li, M., & Zou, B. (2019). A Study on Idiom Translation from the Perspective of Cultural Differences. *Journal of Language Teaching and Research*, 10(1), 182-188.

Li, S., Chen, J., Yuan, S., Wu, X., Yang, H., Tao, S., Xiao, Y., & Shanghai. (n.d.). *Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models*. Retrieved November 16, 2023, from https://arxiv.org/pdf/2308.13961.pdf

Martinez, R., & Schmitt, N. (2020). The effect of frequency and L1 congruency in idiomatic expressions for L2 idiom learning. *Language Teaching Research*, 24(1), 60-79.

Miller, J. (2019) 'The bottom line: Are idioms used in English academic speech and writing?', *Journal of English for Academic Purposes*, 43 (2020) 100810. Available online at: https://doi.org/10.1016/j.jeap.2019.100810.'

Nie, Y., Chen, N. F., & Bansal, M. (2020). Just Add Functions: A Neural Multi-hop Reasoning Framework for Task-Oriented Dialogues. *arXiv preprint arXiv:2005.11016*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Fiederer, R. & O'brien, S. (2009). Quality and Machine Translation: A realistic objective?. The Journal of Specialised Translation 11, 52-74

Schwenk, H., Douze, M., & Barrault, L. (2019). A large-scale multilingual corpus for sentence meaning representation. *arXiv preprint arXiv:1912.01574*.

Simpson, R., and Mendis, D. (2003) 'A corpus-based study of idioms in academic speech', *Tesol Quarterly*, 37(3), 419e441. Available online at: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/90255/3588398.pdf?sequence=1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. ArXiv.org. https://arxiv.org/abs/1706.03762

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019). *Learning Deep Transformer Models for Machine Translation*. https://doi.org/10.18653/v1/p19-1176

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, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., & Stevens, K. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. ArXiv.org. https://arxiv.org/abs/1609.08144

Zhao, Y., Zhang, M., Chen, X., Deng, Y., Geng, A., Liu, L., ... & Zhang, Z. (2023). Human Evaluation for Translation Quality of ChatGPT: A Preliminary Study. *HiT-IT 2023*, 282-287.

Zulfiqar, S., Wahab, A., Sarwar, M., & Ingo Lieberwirth. (2018). *Is Machine Translation a Reliable Tool for Reading German Scientific Databases and Research Articles? 58*(11), 2214–2223. https://doi.org/10.1021/acs.jcim.8b00534