

Designing AI-powered translation education tools: a framework for parallel sentence generation using SauLTC and LLMs

Moneerh Aleedy^{1,2}, Fatma Alshihri³, Souham Meshoul¹,
Maha Al-Harathi⁴, Salwa Alramlawi⁵, Badr Aldaihani⁶, Hadil Shaiba⁵
and Eric Atwell²

¹ Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

² School of Computer Science, University of Leeds, Leeds, United Kingdom

³ Department of Translation, College of Languages, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

⁴ Department of Applied Linguistics, College of Languages, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

⁵ Department of Computer Science, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

⁶ Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

ABSTRACT

Translation education (TE) demands significant effort from educators due to its labor-intensive nature. Developing computational tools powered by artificial intelligence (AI) can alleviate this burden by automating repetitive tasks, allowing instructors to focus on higher-level pedagogical aspects of translation. This integration of AI has the potential to significantly enhance the efficiency and effectiveness of translation education. The development of effective AI-based tools for TE is hampered by a lack of high-quality, comprehensive datasets tailored to this specific need, especially for Arabic. While the Saudi Learner Translation *Corpus* (SauLTC), a unidirectional English-to-Arabic parallel *corpus*, constitutes a valuable resource, its current format is inadequate for generating the parallel sentences required for a didactic translation *corpus*. This article proposes leveraging large language models like the Generative Pre-trained Transformer (GPT) to transform SauLTC into a parallel sentence *corpus*. Using cosine similarity and human evaluation, we assessed the quality of the generated parallel sentences, achieving promising results with an 85.2% similarity score using Language-agnostic BERT Sentence Embedding (LaBSE) in conjunction with GPT, outperforming other investigated embedding models. The results demonstrate the potential of AI to address critical dataset challenges in quest of effective data driven solutions to support translation education.

Submitted 1 October 2024

Accepted 6 March 2025

Published 31 March 2025

Corresponding author

Fatma Alshihri,
faalshihri@pnu.edu.sa

Academic editor

Bilal Alatas

Additional Information and
Declarations can be found on
page 23

DOI 10.7717/peerj-cs.2788

© Copyright

2025 Aleedy et al.

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Artificial Intelligence, Computational Linguistics, Computer Education, Data Science

Keywords Didactic *corpus*, *Corpus* annotation, AI-powered translation education, AI-based translation technology

INTRODUCTION

As the world becomes increasingly interconnected, often described as a “global village”, translation emerges as a pivotal practice facilitating communication and understanding

across diverse languages. Translation bridges cultural differences and plays a vital role in fostering social and cultural development (Zhao, Li & Tian, 2020). The domain of translation education (TE) is central to preparing future translators and interpreters, equipping them with the necessary skills to proficiently translate written and spoken content across linguistic systems. This critical field shoulders the responsibility of enabling global communication. However, TE is inherently dynamic and continuously evolving to keep pace with advancements in translation practices and pedagogical approaches (Igalada & Echeverri, 2019).

Academic translation programs are shaped by two primary forces: the evolving demands of the translation market, driven by technological innovations, and progressive trends in higher education teaching methodologies (Sawyer, Austermühl & Raído, 2019). Recent technological advancements, such as artificial intelligence (AI), big data, and deep learning, have revolutionized the translation industry (Zhao, Li & Tian, 2020; Al-Batineh & Al Tenaijy, 2024). Concurrently, higher education has experienced transformative shifts, including the shift from teacher-centered to student-centered learning models in the late 1990s (Igalada & Echeverri, 2019) and the widespread adoption of competency-based training at the start of the 21st century (Albir, 2017).

To remain relevant, translation programs must adapt their curricula to integrate these technological advancements and align with modern pedagogical practices. This requires the incorporation of state-of-the-art translation tools and the adoption of learning approaches that reflect market needs (Sawyer, Austermühl & Raído, 2019). In response, scholars have underscored the importance of improving existing teaching tools and developing new resources to enhance TE (Zapata, 2016). Key recommendations include revising curricular designs and refining assessment methods to better reflect real-world translation practices (Albir, 2017; He, 2021).

The PACTE group believes that a pivotal aspect of this transformation is the emphasis on translation competence, grounded in cognitive-constructivist and socio-constructivist learning theories (Albir, 2017). The authors in He (2021) advocates for student-led teaching approaches that leverage technological resources. By encouraging students to independently engage with multiple platforms, revise translations, and conduct peer assessments, TE can better reflect the demands of the contemporary translation landscape. Considering emerging technologies, He (2021) emphasizes that translators' primary role will increasingly involve post-editing machine-generated translations, highlighting the need to cultivate post-editing competencies within TE programs.

The advent of AI, particularly in the domains of natural language processing (NLP) and natural language understanding (NLU), has introduced sophisticated tools capable of supporting these objectives. AI-powered applications can personalize learning experiences, analyze student translations, and automate certain assessment processes, thereby enhancing efficiency and objectivity. Such innovations ensure that students remain competitive in the evolving translation market by acquiring the necessary skills to collaborate with AI tools rather than being replaced by them.

Despite these advancements, a significant challenge persists: the scarcity of high-quality, comprehensive datasets tailored to TE. A parallel *corpus*, in this context, is defined as a

collection of texts in two languages that are direct translations of each other. The Saudi Learner Translation *Corpus* (SauLTC) represents a notable attempt to address this gap. Developed to facilitate the teaching of translation from English to Arabic (*Al-Harhi & Al-Saif, 2019*), SauLTC is a multi-version, English-Arabic parallel *corpus* featuring part-of-speech tagging. It provides two target versions of 366 source texts (STs), allowing for the analysis of translation and revision processes from initial to final drafts. This *corpus* supports error identification, the evaluation of teaching feedback, and the investigation of individual differences in manual verification. The SauLTC project further aims to develop AI-based translation technologies and eLearning resources to support digital TE.

In alignment with the evolving landscape of TE, this research seeks to build upon SauLTC by developing a parallel didactic *corpus* enriched with an innovative annotation scheme. A key component of this process is parallel sentence generation, which refers to the automated process of identifying or constructing sentence pairs that convey the same meaning across bilingual texts. Additionally, alignment is the process of matching sentences from two different languages, where each sentence is a translation of the other. This scheme aims to enhance students' translation competence, foster self-directed learning, and promote autonomy (F Alshihri, M Alharhi, 2024, unpublished data). By focusing on linguistic equivalence between English and Arabic, the annotation highlights structural and lexical correspondences, encouraging learners to reflect on key aspects of the translation process, such as translation units, problem-solving strategies, and the impact of errors.

Moreover, the proposed tool incorporates error analysis and post-editing features, allowing students to identify and rectify inaccuracies systematically. This hands-on approach equips students with the practical skills needed to address real-world translation challenges, bridging the gap between theoretical knowledge and applied practice.

The first stage of this project outlines a framework for transforming SauLTC's parallel texts into parallel sentences, laying the foundation for future annotations that stimulate critical thinking and active engagement with translation problems. By harnessing AI and large language models (LLMs), the framework facilitates the generation of high-quality parallel sentences, ensuring contextual appropriateness and linguistic precision. The integration of AI-driven tools enhances the pedagogical value of the *corpus*, offering students a more comprehensive learning experience.

Ultimately, the development of a SauLTC-based parallel sentence dataset promises to advance AI-powered TE tools, providing students with invaluable resources for practical analysis, comparative evaluation, and cultural competency. This initiative supports instructors in delivering targeted feedback, fostering error correction, and enhancing classroom engagement.

The remainder of this article is organized as follows: "Translation and Technology" reviews the role of technology in translation education, while "Related Work" explores recent initiatives in *corpus* development for TE. "Methodology and Proposed Parallel Sentence Generation Framework" describes the methodology and framework for parallel sentence generation. "Experiment and Results" presents the results and analysis, and "Discussion" concludes with future research directions.

TRANSLATION AND TECHNOLOGY

The rapid advancement of information technology, notably AI technology, is one of the realities affecting translation services and teaching/training (*Garbovskiy & Kostikova, 2020; Kanglang & Afzaal, 2021; Wang, 2023; Mohamed et al., 2024; Mohsen, 2024*), resulting in a process of constant evolution in the field (*Sawyer, Austermühl & Raído, 2019; Kenny, 2019*). The translation industry is now AI-driven which improved the levels of machine translation accuracy and efficiency (*Zaghloul & Khasawneh, 2024*). Machine translation tools and CAT tools are constantly updated and upgraded because of the advances in artificial intelligence technologies (*Steigerwald et al., 2022; Mohamed et al., 2024*), and this means that the translator, who is expected to translate a huge number of texts quickly in a little amount of time (*Rodríguez De Céspedes, 2019*), needs to be kept abreast of the latest translation technologies, and therefore, these technologies should be a core component in any TE (*Kenny, 2019*). In fact, translators will not have a place in today's translation market if they do not know how to use translation technology (*Bowker, 2023*).

Technological innovations have led to huge transformations in the translation profession, education, and assessment (*Kong, 2022*). One of the most prominent transformations is the Computer-Assisted translation (CAT) change of the translators' role into post-editors (*Igualada & Echeverri, 2019*). Therefore, translation students need to be trained to use the most recent AI-based translation technologies to enhance their learning experience and empower them to meet the rising demands of the market (*Zaghloul & Khasawneh, 2024*). Embracing AI technologies provides teachers with ample opportunities to create high-quality translation environments with favorable conditions to improve students' foreign language skills, improve their translation skills and broaden their horizons, and support active learning (*Kong, 2022; Liu, 2022; Wang, 2023*).

RELATED WORK

Learner translation *corpora* (LTC) are comprised of translations produced by students, either into their first language (L1) or a second language (L2), aligned with their STs. This subset represents a specialized area within *corpus*-based translation studies (CBTS), which typically focuses on professional or expert translation *corpora* (*Lefer, 2020*). The emergence of LTC began in the early 2000s, approximately a decade following the establishment of CBTS. Foundational initiatives such as PELCRA (*Uzar & Waliński, 2001*) and MeLLANGE (*Castagnoli et al., 2011*) were instrumental in shaping and advancing LTC. *Espunya (2014)* highlights the primary objectives of LTC, which include analyzing the development of translation skills, assessing the efficacy of pedagogical approaches, and developing resources for translator training. Similarly, *Andrey & Kunilovskaya (2014)* present a comprehensive research plan based on their work with the Russian Learner Translator *Corpus* (RusLTC), focusing on translation variability, decision-making processes, and the identification of challenging linguistic areas.

Drawing from the MISTiC *corpus*, *Castagnoli (2020)* observes that full lexical consistency is most evident in the translation of concrete nouns, functional words, and

numerical data, while abstract nouns and metaphorical expressions frequently exhibit greater variability. A more recent contribution to the field of language and translation education is the Multilingual Student Translation *Corpus* (MUST) which is a multilingual LTC project that provides extensive metadata for translation research ([Granger & Lefer, 2020a, 2020b](#)).

Moreover, as part of the MUST initiative, the translation-oriented annotation system (TAS) was developed to enhance translator training and facilitate research on translation quality across various language pairs ([Granger & Lefer, 2021](#)).

The multifunctionality of LTC is evident in its diverse applications. These *corpora* are integral to translator education, offering students exposure to parallel sentence structures that deepen their understanding of linguistic patterns and cultural nuances. They also serve as valuable resources for research on translation quality by facilitating comparative studies of various translation strategies. Furthermore, LTC supports the development of machine translation (MT) systems and computational linguistic studies, underscoring its relevance across both academic and professional domains.

Annotation systems employed in LTC play a critical role in enhancing the analytical process of student translations. Commonly used annotation methods include part of speech (POS) tagging, which labels each word with its grammatical category, allowing for the examination of syntactic patterns and learner language use. Parsing, another key method, analyzes sentence structures to identify syntactic dependencies and relational hierarchies. Error annotation, a cornerstone of LTC research, involves marking errors within learner translations to pinpoint areas of improvement, such as grammatical inconsistencies or lexical inaccuracies.

Several taxonomies within LTC vary widely, ranging from simple frameworks comprising five categories to more elaborate schemes encompassing over 50 classifications. While projects such as RusLTC and KOPE have developed comprehensive taxonomies for English, French, and German pairs, some annotation systems lack sufficient documentation, posing challenges for achieving consistency in error categorization. Nevertheless, error annotation yields valuable insights into translation competence development. For instance, [Wurm \(2020\)](#) leverages KOPE error data to investigate the empirical dimensions of translation proficiency, analyzing factors such as time spent abroad, media exposure, and temporal changes in student performance. Her findings indicate that intensive training significantly reduces errors, enhances problem-solving efficiency, and accelerates translation speeds, whereas external variables such as international experience exhibit limited influence on the acquisition of translation competence.

An emerging trend within CBTS is the annotation of *corpora* focusing on student post editing. Researchers, including [Kübler, Mestivier & Pecman \(2022\)](#), have focused on identifying errors in editing, particularly those involving complex noun phrases in specialized discourse. Their taxonomy categorizes errors into three groups; overconfidence in MT where incorrect MT output remains unchanged; underconfidence in MT, where correct MT output is unnecessarily altered; and failure to correct MT, where errors in MT

are acknowledged but not adequately corrected. In their *corpus*, the error type concerning noun phrases is failure to correct MT.

Further advancing this domain, [Lefer, Piette & Bodart \(2022\)](#) introduced the Machine Translation Post-Editing Annotation System (MTPEAS). It consists of seven categories; value adding edits, successful edits, unnecessary edits, incomplete edits, error introducing edits, unsuccessful edits and missing edits. They have effectively integrated this system with the TAS developed by [Granger & Lefer \(2021\)](#). This classification system helps in structuring annotations for student editing datasets, allowing for the identification and clarification of incorrect segments that persist in the final edited texts. This approach enhances the investigation and understanding of errors encountered during student editing tasks.

These initiatives collectively enrich the scope of LTC research, fostering advancements in translation education and contributing to the diversification of *corpus*-based translation studies. By accommodating multiple languages, text genres, and proficiency levels, LTC continues to serve as a vital resource for advancing translation pedagogy and practice.

SauLTC represents a unidirectional, multi-version parallel LTC, comprising student translators' graduation projects, some of which have undergone professional editing and subsequent publication. SauLTC offers two distinct translation versions: a sub-*corpus* of unedited draft translations reflecting the initial outputs of student translators, and a final sub-*corpus* containing professionally edited versions. The SauLTC software facilitates both independent searches within each sub-*corpus* and parallel searches across two or three sub-*corpora*. The availability of draft and final submission sub-*corpora* allows for the differentiation between linguistic features attributable to the initial translation process and those arising from post-feedback revisions. By analyzing both draft and final versions, researchers can gain deeper insight into the translation process, offering valuable contributions to the understanding of translation development and competence ([Kruger, 2012](#); [Bisiada, 2017](#)). This dual-layer structure enables the identification of mediations occurring during the transition from source to target text and through subsequent editorial interventions. These mediations may manifest as "translationese" (unusual features characteristic of translated texts) or as editorial modifications aimed at enhancing lexical diversity, simplification, or explicitation. The dual availability of draft and final versions positions SauLTC as an essential resource for process-oriented, *corpus*-based translation studies. The authors in [Alasmri & Kruger \(2018\)](#) highlighted the scarcity of *corpus*-based translation Arabic research. Thus, to address this scarcity, particularly in learner *corpus* research and *corpus*-based translation studies, this article works towards a more accurate sentence-alignment system for SauLTC.

Sentence alignment methods fall into three main categories: length-based, lexical-based, and neural approaches. Length-based models ([Church et al., 1993](#)) align sentences based on length correlation, which has been effective for some European languages but struggle with morphologically rich languages and those with different character structures. Lexical-based methods ([Varga et al., 2007](#)) align sentences using bilingual dictionaries and word-overlap heuristics, making them effective for direct translation equivalents. However, they struggle with structurally divergent translations and

paraphrased sentences, where direct word matches are insufficient for accurate alignment. Authors in [Grégoire & Langlais \(2018\)](#) introduced a neural-based parallel sentence extraction system that maps sentences into a shared vector space, improving alignment accuracy over traditional methods. Their fully neural approach eliminates the need for predefined feature engineering but relies on large-scale parallel *corpora*, making it less applicable to low-resource languages. Additionally, their model does not handle out-of-vocabulary (OOV) words, which may lead to alignment errors when dealing with rare or unseen terms. The Parallel Hierarchical Attention Network (PHAN) ([Zhu, Yang & Xu, 2020](#)) further improves alignment by assigning different weights to key words, yet it also depends on large parallel datasets for training, posing challenges for low-resource language pairs.

METHODOLOGY AND PROPOSED PARALLEL SENTENCE GENERATION FRAMEWORK

To address the problem at hand, this section details the methodological framework employed in this study to derive a didactic *corpus* of parallel sentences from the SauLTC. The chosen approach ensures a robust and systematic investigation of recent LLMs in achieving parallel sentence generation from document pairs. The following sub-sections dive into the specific methods utilized to achieve the research objectives. [Figure 1](#) shows the end-to-end methodological framework adopted in our study. It encompasses the following main phases:

Data collection and understanding

As stated earlier, we use the SauLTC developed as part of the “Design and Compilation of the Saudi Bilingual *Corpus* for Translation Learners” project ([Al-Harathi & Al-Saif, 2019](#)). SauLTC is an English-Arabic translation *corpus* comprising three sub-*corpora* and three participant profiles. The first *corpus* consists of English student STs; the second and third *corpora* include two versions of the translations into Modern Standard Arabic (MSA).

The SauLTC project involved three types of participants. The first and primary group consisted of 366 senior Saudi female translation students, with a mean age of 21.84 years (standard deviation (SD) = 0.64). These students were in the final stages of their academic program, ensuring a consistent level of translation proficiency for the study. The student profile contains information about the student translator, the source, and the translated text. The second group of participants included 48 women instructors who provided feedback on students’ draft translations and later assessed their final submissions. All instructors hold at least a master’s degree in translation or linguistics and with diverse employment backgrounds. Some instructors were college faculty members, while others were freelancers or loan faculty members from other universities in Saudi Arabia (see [Al-Harathi et al., 2024](#)) for more details. Instructor profiles document their educational background, teaching experience, and expertise in supervising translation projects. The third group of participants was composed of 23 alignment verifiers who reviewed and verified the accuracy of the automatic sentence alignments.

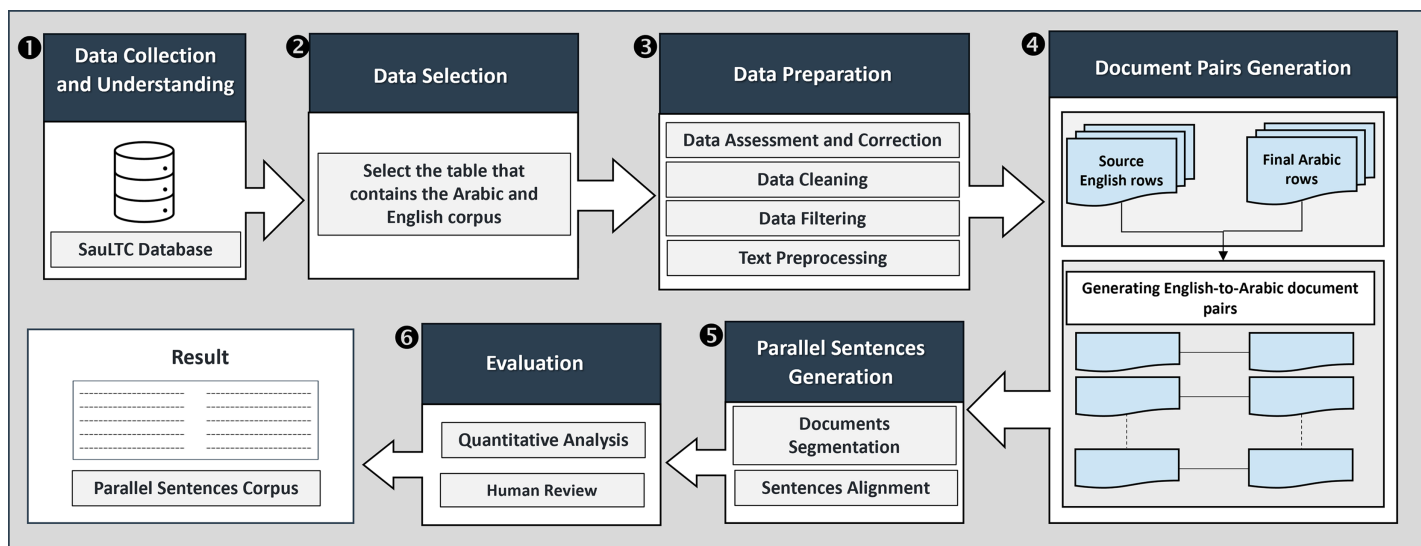


Figure 1 Proposed methodological framework.

Full-size DOI: 10.7717/peerj-cs.2788/fig-1

The STs used in the SauLTC *corpus* are chapters or extracts from booklets, averaging approximately 6,000 words each. Students had the option to select their own texts for translation, provided they obtained instructor approval for the book title before beginning the translation process. These texts cover a variety of genres, as detailed in Table 1, including psychology, self-help, business, parenting, language, religion, education, biography, autobiography, nutrition, management, fiction, social sciences, sciences, history, and health-related texts. To address potential overlaps between genres like “business” and “management,” we defined them based on scope and focus. “Business” covers commercial activities, market strategies, and economic interactions, while “management” focuses on organizational processes, leadership, and decision-making. Classification was guided by keyword analysis, content review, and purpose-based categorization. Ambiguous cases were resolved by prioritizing the text’s primary focus, ensuring clear distinctions for data interpretation. Table 1 lists the text genres available in the database (Al-Harathi & Al-Saif, 2019).

Data selection

This section outlines our approach to selecting the essential parts of the SauLTC for our translation learning system. The dataset was carefully reviewed to identify the most relevant components for our study. We specifically focused on the dataset that contains English texts alongside their corresponding Arabic translations. This targeted selection from the dataset ensures that our analysis is precisely aligned with the needs of bilingual translation studies. Table 2 below shows the description of the related metadata.

We targeted health-related texts for our analysis as they represent the largest subset of translations in the *corpus*, comprising 109 texts. This substantial subset allowed for a detailed exploration of their linguistic and translational features while reflecting the predominant composition of the *corpus*. However, the framework used in this study is designed to work equally well with all text genres, requiring no specific adjustments or

Table 1 SauLTC text genres.

Text genres	Number of texts in the database under each text genre
Health	109
Psychology	41
Self-help	66
Business	51
Parenting	12
Language	12
Religion	1
Education	28
Biography	1
Autobiography	20
Nutrition	11
Management	1
Fiction	7
Social sciences	1
Sciences	3
History	2
Total	366

Table 2 Dataset metadata.

Column name	Description
ID	Row sequence number
Type	Type of translation (SOURCE, DRAFT, FINAL)
groupNumber	Translation file group number
transID	Translator ID
Version	Translation version
Year	Translation submission year
Text	The English or Arabic text
Genre	The classification of the text

handling for different genres. This ensures that the findings are applicable across various genres within the SauLTC and that the alignment process remains consistent regardless of genre.

Data preparation

The data preparation phase is critical in developing LLMs, especially when dealing with the complexities of English-Arabic texts. This phase involves several key steps to ensure the dataset is optimized for subsequent analyses:

- **Initial data assessment and correction:** We began with a comprehensive review of the dataset to identify and document any discrepancies or mismatches. This included

checking for instances where English STs did not have corresponding Arabic translations and vice versa. We also addressed and corrected any data entry errors, particularly where texts are mislabeled or placed in incorrect database columns. Ensuring that each text is correctly categorized is essential for accurate data handling and subsequent processes.

- **Data cleaning:** We meticulously cleaned the dataset by identifying and eliminating non-essential columns such as ID, group number, version, and year. This step simplifies the dataset and focuses attention on the core aspects of the translation *corpus*.
- **Data filtering:** We filtered out all entries containing preliminary draft texts, opting to retain only those entries that included the final revised versions of the Arabic translations corresponding to their English STs. This selective inclusion ensures that our analysis is based on the most accurate and refined data available.
- **Text preprocessing:** The text preprocessing phase involved carefully planned steps to preserve the integrity and meaning of the text while addressing formatting irregularities. Regular expressions were used to handle issues such as extraneous spaces, newline characters (`\n`), line breaks (`\n`), and non-breaking spaces (`\xa0`), ensuring a clean and standardized text format. This approach maintained the original structure and semantics, making the data suitable for alignment and analysis. [Figure 2](#) shows a general example of text preprocessing process.

Document pairs generation

Once the dataset was cleaned and verified, advanced scripting and database management tools were employed to automate the pairing of English STs with their corresponding Arabic translations. This process ensured the accuracy and efficiency of generating document pairs.

Parallel sentences generation

Following the successful pairing of text documents, the next pivotal step in our methodology involves the generation of parallel sentences pairs in English and Arabic. For this task, we employed the Generative Pre-trained Transformer (GPT) model, developed by OpenAI. GPT leverages transformer-based architecture and serves as a prime example of a LLM. Its versions GPT-3 and GPT-4 are some of the most well-known and powerful LLMs currently available. This model is trained on a diverse *corpus* of text data and excels in recognizing and generating sentence boundaries, not merely based on punctuation but through a deep contextual understanding of language syntax and structure. GPT's sophisticated capabilities allow it to handle nuanced expressions and complex grammatical constructions effectively, surpassing traditional methods in segmenting and aligning sentences in complex texts ([Wu et al., 2023](#)).

To facilitate sentence alignment using this LLM through GPT, we implemented a multi-step methodology to address some limitations of the GPT model. Initially, we utilized a Sentence Embedding model to segment the texts into smaller units based on semantic similarity and a predefined token count limit for English.

Sentences Preprocessing Example

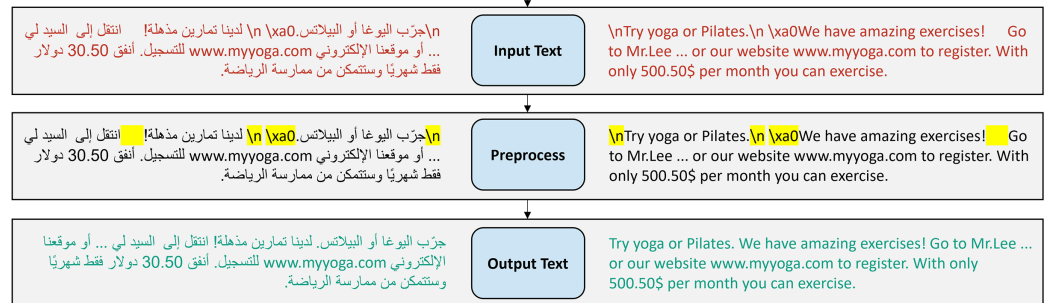


Figure 2 Preprocessing examples.

Full-size DOI: 10.7717/peerj-cs.2788/fig-2

The segmentation process began by splitting the English text into smaller units, each limited to approximately 1,000 tokens. This process employed sentence-level tokenization to segment the text into coherent, semantically meaningful units. Sentence boundaries were identified using basic delimiters, such as punctuation and whitespace, ensuring sentence structure was preserved while adhering to the token limit.

For Arabic, corresponding segments were identified dynamically using multilingual embeddings to ensure semantic equivalence with the English segments. Specifically, the cosine similarity of sentence embeddings was calculated to match the last sentence in each English segment with the most semantically similar Arabic sentences. This dynamic alignment determined the Arabic segment boundaries, ensuring contextual and semantic alignment with the English segments.

Although the Arabic segments were not explicitly constrained by a fixed token limit, the semantic alignment process ensured they were approximately equivalent in length to the English segments in terms of tokens. The underlying Sentence Embedding model internally employs subword-level tokenization, enabling accurate representation of linguistic nuances in both English and Arabic, including morphological variations and structural differences.

By integrating sentence-level segmentation for English and semantic similarity-based alignment for Arabic, this approach ensures that the English and Arabic segments are balanced in terms of tokens and content, making them suitable for subsequent parallel sentence generation.

For sentence alignments, each segment pair is processed by GPT to split the segments into aligned sentences. The following prompt was used to instruct GPT:

- Prompt: “Segment each English text and Arabic text, then respond by aligning each sentence in the English text with its corresponding Arabic text. Format each pair as ‘English: (English Text) Arabic: (Arabic Text)’. Separate each sentence pair with a newline.”

The responses from GPT are long texts where each English text segment is introduced with the keyword “English:” followed by its corresponding Arabic translation, which starts with the keyword “Arabic:”.

Despite GPT’s advanced capabilities, it may occasionally encounter technical and formatting issues, necessitating a thorough review and correction process to ensure accuracy. In this study, we assess the quality of GPT’s responses and make necessary adjustments to align with our quality criteria. These criteria include ensuring there are no null values, and that all data adheres to the correct format. Responses that meet these standards are archived for future use, while discrepancies are carefully documented for in-depth analysis.

The detailed procedure for generating parallel sentences is outlined in the pseudocode illustrated in Fig. 3.

Evaluation

The evaluation phase of our study plays a critical role in assessing the effectiveness and accuracy of the parallel sentence pairs generated in the previous stages. This phase utilized a dual-method approach: quantitative analysis through advanced multilingual sentence embedding models and qualitative analysis *via* expert human review.

The quantitative analysis involved transforming sentences into numerical vectors using several multilingual sentence embedding models. This conversion allows for a robust semantic alignment analysis between the English and Arabic sentences by calculating cosine similarity scores. The proposed model in this research includes: the Universal Sentence Encoder (mUSE) (Reimers & Gurevych, 2020; Feng et al., 2020), MiniLM (Kurek et al., 2024), MPNet (Kurek et al., 2024) and the Language-agnostic BERT Sentence Embedding (LaBSE) (Reimers & Gurevych, 2020; Feng et al., 2020). These proposed models were selected for their:

- **Efficacy in handling multilingual content:** These models are trained on massive datasets encompassing multiple languages, enabling them to understand the nuances of different languages and produce comparable embeddings.
- **Ability to capture nuanced linguistic features:** By leveraging powerful architectures like transformers, these models go beyond simple word co-occurrence and capture deeper semantic relationships within sentences.

Once the text segments are transformed into embeddings, we employ cosine similarity to measure the semantic similarity between each pair of English and Arabic vectors. Cosine similarity is a metric used to determine how similar two vectors are, irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity formula, as described by Alfarizy & Mandala (2022), is given in Eq. (1):

$$\text{Cosine Similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} \quad (1)$$

Pseudocode: Bilingual Text Segmentation, Aligning using GPT	
Input:	A prepared data file containing bilingual (English and Arabic) documents.
Output:	Two data files: 1. Semantically aligned text segments. 2. Aligned sentences.
Algorithm	
Part 1: Generate Semantically Aligned Text Segments:	<ul style="list-style-type: none"> • Iterate over each English-Arabic text pair in the data file. • Use a sentence transformer model to segment texts into smaller parts based on semantic similarity and a predefined word limit.
Part 2: Align and Evaluate Sentences Based on Semantic Similarity:	<ul style="list-style-type: none"> • For each pair of semantically aligned segments: <ul style="list-style-type: none"> • Pass the English and Arabic segments to GPT for further alignment. • Retrieve and save GPT responses. • Reformat the aligned sentences into structured table. • Check and refine the alignment based on quality assessments.

Figure 3 Pseudocode of the bilingual text segmentation, aligning using GPT.

Full-size  DOI: [10.7717/peerj-cs.2788/fig-3](https://doi.org/10.7717/peerj-cs.2788/fig-3)

where A and B are the vector representations of the English and Arabic text segments, respectively. The resulting value ranges from -1 to 1 :

- 1 : The texts are exactly similar.
- 0 : The texts have no similarity.
- -1 : The texts are exactly opposite.

To ensure accurate alignment, we established a cosine similarity threshold of 0.7 for identifying semantically aligned sentence pairs. This threshold was chosen based on empirical findings from studies in multilingual sentence alignment tasks. For instance, [Chimoto & Bassett \(2022\)](#) demonstrated that restricting sentence pairs to those with cosine similarity scores above 0.7 yielded alignment accuracies exceeding 85% .

However, the effectiveness of a 0.7 threshold can vary depending on the embedding model used. As discussed in [Song et al. \(2024\)](#), cosine similarity thresholds are influenced by the training methodologies of embedding models, which may produce higher similarity scores even for dissimilar sentences. To address this, the human evaluation process plays a critical role in identifying and addressing such cases where cosine similarity alone may fail to capture nuanced semantic differences.

In the human evaluation process, a representative subset of 100 sentence pairs was selected for review. Random sampling was employed to minimize selection bias and ensure the sample reflected the broader dataset's characteristics. Random selection is a widely recognized method in computational linguistics for maintaining generalizability and avoiding overfitting to specific subsets ([Saldías et al., 2022](#)). A sample size of 100 pairs balances practical feasibility with statistical reliability, as demonstrated by similar practices in machine translation evaluation studies, such as the 100 -sentence test sets used for validation in English Korean translation experiments ([Park & Padó, 2024](#)).

The evaluators involved in the human assessment were bilingual experts with extensive experience in translation, linguistic analysis, and *corpus* linguistics. They were university

faculty members specializing in translation education and *corpus* linguistics, ensuring pedagogical and practical insights into translation quality assessment. Both evaluators had been involved in multiple evaluation processes and underwent a short training phase before commencing evaluation.

Evaluators utilized a standardized rubric inspired by the Multidimensional Quality Metrics (MQM) framework (Freitag et al., 2021; Lommel et al., 2024), and were familiarized with the evaluation criteria, which covered addition, omission, and mispairing errors. The rubric minimizes bias by applying standardized, objective criteria for error identification and scoring, ensuring consistency across evaluations (Freitag et al., 2021; Lommel et al., 2024). Our evaluation focused on three key factors related to accuracy: addition, which refers to extra information added that is not present in the source; omission, which involves information missing that is present in the source; and mispairing, where the incorrect meaning is conveyed compared to the source (Freitag et al., 2021; Lommel et al., 2024).

Each error type was classified into one of three severity levels: minor errors, which do not significantly impact the understanding or usability of the translation; major errors, which affect understanding or usability but can still be inferred; and critical errors, which severely impact understanding or usability and may cause significant misunderstandings. These severity levels were assigned specific weights—minor (1), major (5), and critical (25)—to reflect their impact on translation quality (Freitag et al., 2021; Lommel et al., 2024). This classification aligns with other frameworks in translation quality evaluation. For instance, O'Brien (2012) defines minor errors as those that are noticeable but do not impact usability, major errors as those that negatively affect meaning, and critical errors as those that significantly impair usability, safety, or behaviour. These definitions align closely with those outlined by Malcolm (2004), who emphasizes the varying degrees of impact on functionality and safety.

The error types and severities are illustrated in Table 3, which provides examples for each category to clarify their impact on the accuracy of the sentence pair alignment.

The overall MQM quality score was calculated using Eq. (2):

$$\text{MQM Quality Score} = 100 - \left(\frac{\sum_{i=1}^n w_i \cdot e_i}{T} \cdot 100 \right) \quad (2)$$

where w_i represents the weight of each error type, e_i is the frequency of errors, and T is the total evaluated sample.

Errors identified during human evaluation were categorized, weighted, and summed according to MQM guidelines to derive the final quality score (Lommel et al., 2024).

To ensure the consistency and reliability of human judgments, Gwet's AC1 coefficient was computed as a measure of inter-rater reliability. It is recognized for its strengths in providing reliable agreement measures even under challenging rating conditions, such as imbalanced data distributions and small sample sizes (Gwet, 2008; Wongpakaran et al., 2013).

Each evaluator independently assessed the same 100-sentence subset, assigning severity ratings to identified errors. These ratings were then recorded and analyzed to quantify the

Table 3 Examples of error types, evaluation scales, and their impact on sentence alignment quality.

Error type	Evaluation scale	Source (English)	Target (Arabic)	Explanation
Addition	Minor	Take a shower and then dry the body.	خذ حمامًا ثم جفف الجسم تمامًا بمنشفة دافئة.	The addition of “بمنشفة دافئة” (“with a warm towel”) introduces extra, non-essential information but does not significantly alter the meaning.
	Major	Platelets are essential for the body’s ability to control bleeding.	الصفائح الدموية ضرورية لقدرة الجسم على التحكم في الدم ومنع تجلطه.	Adding the phrase “ومنع تجلطه” (stop clotting), which gives inaccurate information, is a major error in the translation, negatively affecting meaning and misleading the audience.
	Critical	Do not mix these chemicals together.	لا تخلط هذه المواد الكيميائية بعضها البعض لأنها آمنة للاستخدام.	Adding “لأنها آمنة للاستخدام” (“because they are safe to use”) misrepresents the warning, creating a critical misunderstanding.
Omission	Minor	The baby is crying loudly.	الطفل يبكي.	The omission of “loudly” (“بصوت عالٍ”) reduces detail but does not affect the primary meaning of the sentence.
	Major	Use disposable needles for one-time use.	استخدام الإبر مرة واحدة فقط.	The omission of “disposable” (“الإبر المخصصة للاستعمال الواحد”) can cause ambiguity, potentially leading to reuse of needles meant for disposal.
	Critical	Examine blood donors for safety and ensure it is free of the AIDS virus.	افحص المتبرعين بالدم.	Omitting “ensure it is free of the AIDS virus” removes vital information about screening for specific, critical risks.
Mispairing	Minor	Take the pill after meals.	تناول الحبة أثناء الأكل.	The target text “أثناء الأكل” (“during meals”) introduces a slight timing discrepancy compared to the source instruction “after meals,” creating a minor mismatch in meaning.
	Major	For external use only.	للاستخدام في الخارج فقط.	The target text “للاستخدام في الخارج فقط” (“To be used outdoors only”) refers to location where the medication should be used, significantly altering the intended meaning of the ST, which instructs users on how the medication should be used and thus creating a major mismatch in meaning.
	Critical	Consult your doctor if symptoms persist.	استمر في تناول الدواء حتى تختفي الأعراض.	The target text “استمر في تناول الدواء حتى تختفي الأعراض” (“continue taking the medication until symptoms disappear”) replaces the source instruction with a completely unrelated statement, misrepresenting the original message.

level of agreement. Gwet’s AC1 was calculated using the probabilistic model (Gwet, 2008), which accounts for the probability of agreement due to chance and adjusts accordingly. The AC1 values were interpreted using the benchmark scale, where values below 0.20 indicate slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and values above 0.81 represent almost perfect agreement (Wongpakaran et al., 2013).

The Gwet’s AC1 Coefficient was calculated using Eq. (3):

$$AC1 = 1 - \left(\frac{D_o - D_e}{1 - D_e} \right) \quad (3)$$

where:

- D_o = Observed disagreement, calculated as the proportion of instances where evaluators provided different ratings.

- D_e = Expected disagreement by chance, determined based on the assumption that ratings are assigned randomly.

In cases where evaluators disagreed significantly, discrepancies were reviewed and discussed to refine judgment criteria. This process helped ensure consistency before final evaluation.

Human evaluators played a crucial role not only in scoring but also in offering qualitative feedback. Their observations highlighted specific alignment challenges and provided deeper context for interpreting results.

EXPERIMENT AND RESULTS

In this section, we detail the experimental setup and procedures used to test and validate the parallel sentence generation framework previously described. This experiment was specifically designed to evaluate the effectiveness of our methodology in generating high-quality English-Arabic sentence pairs with a particular focus on the health text genre within the SauLTC.

Following the completion of the data collection, comprehension, selection, preparation, and document pair generation phases, as elaborated in the preceding section, a total of 85 document pairs were successfully created and prepared for use in the parallel sentence generation phase. These document pairs were subsequently segmented into smaller units to facilitate processing by GPT. The segmentation process involved splitting the English text into smaller units, each limited to approximately 1,000 tokens, using sentence-level tokenization. Arabic segments were dynamically aligned with their corresponding English segments by leveraging semantic similarity scores calculated using the paraphrase-multilingual-mpnet-base-v2 embedding model. This ensured contextual and semantic equivalence between the segments, even though Arabic segments were not constrained by a fixed token limit. The technical details of this process are outlined in [Table 4](#).

The segmented pairs were processed by GPT, configured as described in [Table 4](#), to generate aligned English-Arabic sentence pairs. The experiment utilized the GPT-3.5 Turbo model, chosen for its balance between accessibility, cost-effectiveness, and suitability for the task. The process used a structured prompt to align segments, producing responses with clear formatting. GPT outputs were designed to pair each English segment with its corresponding Arabic translation, marked with the keywords “English:” and “Arabic:” for clarity.

Upon generating the sentence pairs, each response underwent a thorough evaluation to verify its adherence to the quality and formatting standards established for this study. Responses with mismatched English and Arabic counts, improper formatting, or alignment errors were flagged for reprocessing. These flagged outputs were either corrected manually or excluded if they failed to meet the quality criteria upon re-evaluation. During the evaluation process, several issues were identified in the alignment of English-Arabic parallel sentences; examples are presented in [Table 5](#).

These examples illustrate the challenges in ensuring accurate alignment of parallel sentences and demonstrate the importance of our review and correction processes.

Table 4 Technical specifications.

Technical specification	
Embedding model for segment alignment	paraphrase-multilingual-mpnet-base-v2, selected for multilingual capabilities and semantic accuracy in alignment tasks.
Segmentation methodology	Sentence-based segmentation for English (token limit: 1,000); dynamic token-based alignment for Arabic.
Tokenization approach	Sentence-level tokenization for segmentation; subword-level tokenization within embeddings.
GPT model	gpt-3.5-turbo-16k
Prompt	Segment each English text and Arabic texts, then respond by aligning each sentence in the English text with its corresponding Arabic text. Format each pair as: “English: (English Text) Arabic: (Arabic Text)” Separate each sentence pair with a newline.
Number of tokens per message	Approximately 2,000 tokens per message (1,000 for English, dynamically equivalent for Arabic).
Temperature	Set to 0.4 to prioritize deterministic and consistent outputs with minimal randomness in sentence alignment.
Number of messages to GPT	469 total API calls made to process the dataset.

Table 5 Examples of the identified issues in English-Arabic parallel sentence alignment.

Problem	Example
Incomplete	English: Its corresponding Arabic translation is missing. Arabic: (Empty)
Duplication	English: They also just look cool. Arabic: كما أنها تبدو رائعة. English: They also just look cool. Arabic: كما أنها تبدو رائعة.
Language switching	English: الحقيقة هي أن هذا لا يعمل دائما Arabic: The truth is that this doesn't always work.
Repeated Arabic text	English: تذكر أن عملية الهضم تحرق الطاقة! Arabic: !تذكر أن عملية الهضم تحرق الطاقة
Formatting	Building on Enthusiasms English: Building on Enthusiasms Arabic: مبني على الحماس I spoke a bit about this in the chapter on conception. Arabic: تحدثت قليلاً عن هذا في الفصل المتعلق بالحملا. October It's the scariest month of the year, but never fear ... we have tips to get you through the candy, cold weather, and other challenges of October. Arabic: أكتوبر هو أكثر الشهور رعباً في السنة، ولكن لا تخف... لدينا نصائح لمساعدتك في التغلب على الحلوى والطقس البارد والتحديات الأخرى التي تواجهك في أكتوبر.

Of the 469 initial responses generated by GPT model, 443 were accepted as meeting the high standards set for inclusion in the study. The remaining 26 responses did not meet the necessary criteria and were therefore discarded. This process of validation ensures that our data is both reliable and robust, suitable for further analysis and application in translation studies. Ultimately, the results of the sentence alignment processes are organized into a

GPT Response:

English: Step 1 inform yourself If you or a loved one has been diagnosed with cancer, or even if you simply want to do all you can to avoid it, your first step is to understand what cancer is and how you can help your body fight it.

Arabic: الخطوة الأولى تتقف نفسك اذا تم تشخيصك أو تشخيص من تحب بالسرطان أو في خطوتك الأولى ,حال كنت ترغب ببساطة أن تفعل كل ما بمقدورك لتجنب الإصابة به هي أن تفهم ما هو السرطان وكيف يمكنك أن تساعد جسمك على مكافحته.

English: The knowledge you gain here will help you to take action to reduce your risk. If you already have a diagnosis, you will be better able to understand what your doctors tell you, know what questions to ask and then decide, with them, on the choice of treatments to ensure the best possible outcome.

Arabic: وستساعدك هذه المعارف التي ستكتسبها هنا على اتخاذ الإجراءات اللازمة للتقليل أما في حال تم تشخيصك بالمرض فسوف تكون أقدر على أن تفهم ما من خطر الإصابة به يخبرك به أطباؤك وتعرف ماهي الأسئلة التي ينبغي أن تطرحها ومن ثم تقرر معهم على ضوءها خيار العلاج الذي يضمن لك أفضل النتائج الممكنة.



English	Arabic
Step 1 inform yourself If you or a loved one has been diagnosed with cancer, or even if you simply want to do all you can to avoid it, your first step is to understand what cancer is and how you can help your body fight it.	الخطوة الأولى تتقف نفسك اذا تم تشخيصك أو تشخيص من تحب بالسرطان ,أو في حال كنت ترغب ببساطة أن تفعل كل ما بمقدورك لتجنب الإصابة به , فإن خطوتك الأولى هي أن تفهم ما هو السرطان وكيف يمكنك أن تساعد جسمك على مكافحته.
The knowledge you gain here will help you to take action to reduce your risk. If you already have a diagnosis, you will be better able to understand what your doctors tell you, know what questions to ask and then decide, with them, on the choice of treatments to ensure the best possible outcome.	وستساعدك هذه المعارف التي ستكتسبها هنا على اتخاذ الإجراءات اللازمة للتقليل من خطر الإصابة به. أما في حال تم تشخيصك بالمرض فسوف تكون أقدر على أن تفهم ما يخبرك به أطباؤك وتعرف ماهي الأسئلة التي ينبغي أن تطرحها ومن ثم تقرر معهم على ضوءها خيار العلاج الذي يضمن لك أفضل النتائج الممكنة.

Figure 4 Example of converting GPT responses.

Full-size DOI: 10.7717/peerj-cs.2788/fig-4

structured table, formatted with two columns—one for English and one for Arabic, as depicted in Fig. 4.

As a result from the sentence generation phase, we generate a total of 15,845 English-Arabic sentence pairs. The aligned sentence pairs were then evaluated through a dual approach—quantitative analysis and expert human review. Quantitative assessment involved the use of four multilingual sentence embedding models to convert the sentences into vectors for semantic alignment analysis. The models selected for this experiment included:

- “**distiluse-base-multilingual-cased-v2**”: based on the Universal Sentence Encoder (mUSE).
- “**paraphrase-multilingual-MiniLM-L12-v2**”: a smaller and more efficient version of the MiniLM architecture.
- “**paraphrase-multilingual-mpnet-base-v2**”: based on the MPNet.
- “**LaBSE**”: The Language-agnostic BERT Sentence Embedding (LaBSE).

Following the transformation of sentences into vector embeddings using the selected multilingual models, the semantic alignment between English and Arabic sentence pairs was assessed using cosine similarity, a metric that quantifies semantic similarity based on vector relationships. The average cosine similarity scores and SD for each model are presented in Table 6, reflecting their alignment performance and variability.

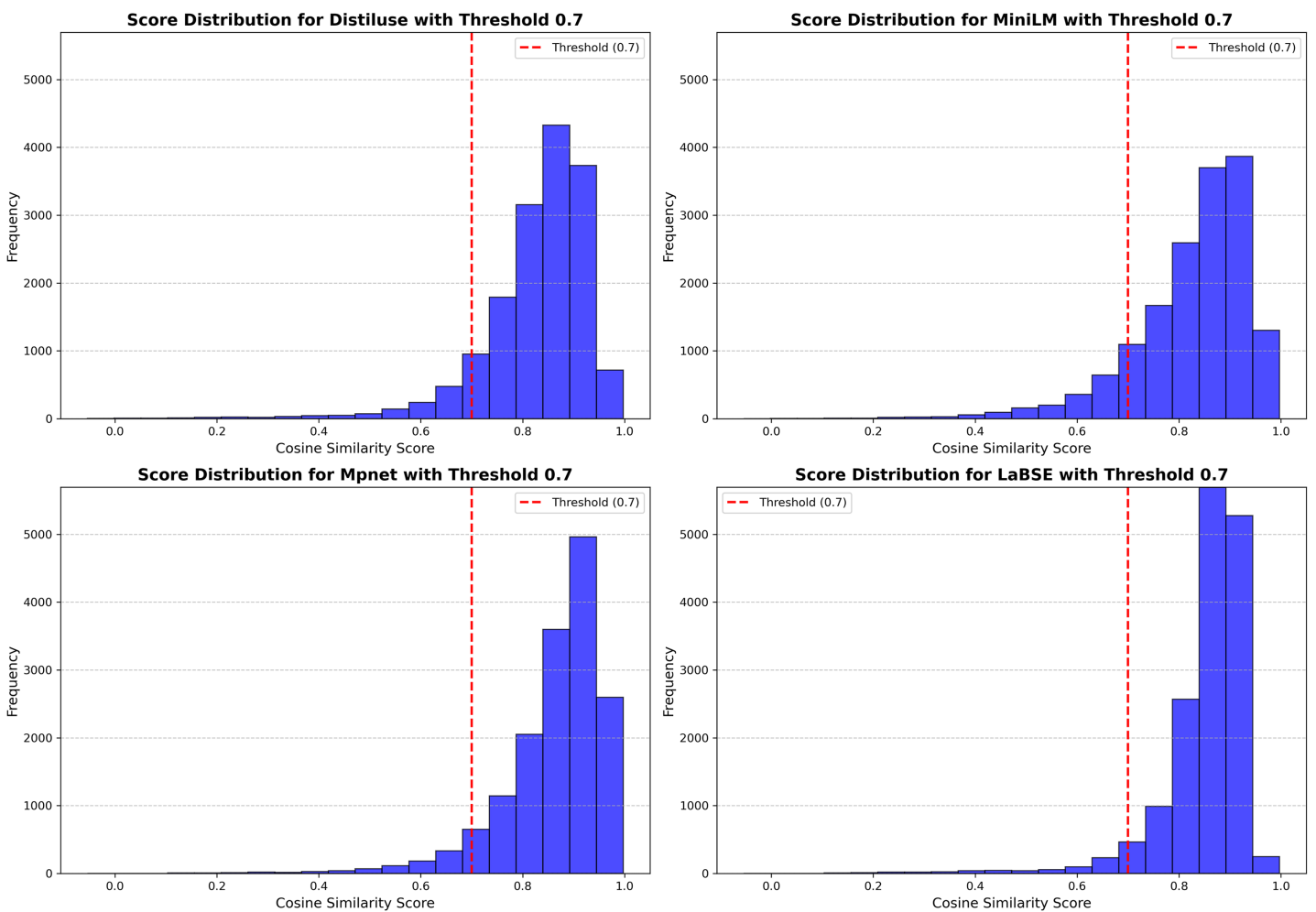
Among the evaluated models, Mpnet achieved the highest average similarity score 0.863, indicating effective semantic alignment. LaBSE followed closely with a score of 0.852 and demonstrated greater consistency, as evidenced by its lower standard deviation (SD = 0.09). These results suggest that while Mpnet provides strong overall alignment, LaBSE offers more consistent performance across diverse sentence pairs.

A cosine similarity threshold of 0.7 was used to categorize sentence pairs:

Table 6 Average cosine similarity scores and SD.

Embedding models	Average cosine similarity scores	SD
Distiluse-base-multilingual-cased-v2	0.828	0.11
Paraphrase-multilingual-MiniLM-L12-v2	0.826	0.12
Paraphrase-multilingual-mpnet-base-v2	0.863	0.10
LaBSE	0.852	0.09

Score Distributions Across Models

**Figure 5** score_distributions_comparison_uniform_frequency.

Full-size DOI: 10.7717/peerj-cs.2788/fig-5

- Aligned pairs: Scores ≥ 0.7 were considered semantically aligned.
- Misaligned pairs: Scores < 0.7 indicated misalignments.

Figure 5 illustrates the distribution of cosine similarity scores, with the threshold prominently marked.

At the 0.7 threshold, LaBSE aligned the highest percentage of sentence pairs (95%), followed by Mpnet (94%), Distiluse (91%), and MiniLM (88%). While these percentages provide a broad overview, examining specific examples shows more about each model's performance.

For instance, the English sentence "Instead, your body will try to keep you alive by slowing down its metabolism" and its Arabic equivalent "بدلاً من ذلك، سيحاول جسمك الحفاظ على حياتك من خلال إبطاء عملية الأيض" were scored differently across models. LaBSE achieved the highest score (0.919), closely followed by Mpnet (0.841), demonstrating their ability to align nuanced sentence pairs effectively. In contrast, Distiluse (0.656) and MiniLM (0.646) struggled to capture the full semantic equivalence, scoring below the 0.7 threshold.

This example aligns with the overall percentages, indicating that LaBSE and Mpnet are more suitable for tasks requiring precise semantic alignment.

After calculating the cosine similarity scores, a comprehensive human evaluation was conducted to further validate the quality of the aligned sentence pairs. A subset of 100 sentence pairs was evaluated by human experts based on predefined criteria outlined in the Methodology section. The human evaluators assigned an average MQM quality score of 85%, indicating a high level of satisfaction with the results. To ensure the consistency and reliability of human judgments, Gwet's AC1 Coefficient was computed as a measure of inter-rater reliability, yielding a score of 0.98, which indicates almost perfect agreement.

A key observation was that sentence pairs with high cosine similarity scores (>0.70) generally aligned well with human evaluations, confirming the reliability of cosine similarity as an initial filtering metric. However, cases involving idiomatic expressions or culturally adapted translations sometimes received lower similarity scores due to differences in surface-level wording. For example, the English phrase "We want what we are told we can't have" was translated as "فكل ممنوع مرغوباً", receiving a similarity score of less than 0.5 across all models. Despite the low score, human evaluators recognized it as a valid translation, illustrating the limitations of cosine similarity in capturing deeper semantic relationships.

Overall, our study confirms that cosine similarity serves as a useful complementary tool in translation evaluation, particularly for identifying lexical alignment. However, human judgment remains essential for assessing meaning in cases involving idiomatic expressions or significant paraphrasing.

DISCUSSION

This study introduces a novel approach to parallel sentence alignment by integrating GPT, advanced sentence embedding models and human expert reviews into a single, adaptive framework.

Traditional length-based and lexical-based methods rely on fixed heuristics or bilingual dictionaries, which struggle with paraphrased translations and structurally divergent sentences. Similarly, neural models improve alignment by mapping sentences into shared vector spaces, but they typically require large parallel datasets for training, making them less suitable for low-resource languages. Many of these models also struggle with out-of-vocabulary words, affecting alignment accuracy when dealing with rare or unseen terms.

Unlike previous models, our approach reduces the need for predefined segmentation rules, as GPT utilizes contextual understanding to assist in identifying sentence boundaries. Our method effectively handles OOV words, paraphrased sentences, and structural variations without requiring extensive *corpus*-specific tuning. Additionally, our framework remains adaptable and scalable across different languages and text genres, overcoming the dataset limitations of earlier models.

Our approach combines automated alignment with expert validation, ensuring both precision and adaptability, particularly LTC. By addressing key challenges in previous methods, it offers a more flexible and efficient solution for parallel sentence generation.

We identified several challenges that impacted the accuracy of the alignment process. One significant issue was inconsistency in sentence length and structure between English and Arabic. This inconsistency is primarily due to some translations being shortened, generally translated, or not matching the source text. For example, the English word “Yes” is translated into Arabic as “نعم، بالتأكيد” which is longer because the Arabic version adds emphasis and certainty with “بالتأكيد” (certainly). This problem is more commonly observed in translating English words that are not lexicalized in Arabic. For example, the English sentence “Metabolism is a word frequently used and seldom understood by dieters” is translated into Arabic as “الأيض هي كلمة كثيرًا ما تستخدم، ونادرًا ما يفهمها الأشخاص الذين يتبعون النظام الغذائي.” The Arabic translation explicitly expands “dieters” to “الأشخاص الذين يتبعون النظام الغذائي” (people who follow a diet), because Arabic does not have one lexical item to express the meaning of the English term “dieters”. In general, these differences arise due to variations in grammar, cultural preferences, and contextual expression.

Another specific challenge arose from instances where students added acknowledgments or extra information at the beginning of the Arabic texts, which were not present in the English versions. This inclusion created additional content in the Arabic translations that had no corresponding segments in the English texts, complicating the creation of accurate translation pairs.

Furthermore, biases inherent in GPT influenced the alignment process and introduced several challenges that required manual interventions to maintain alignment quality and consistency. These challenges included issues such as incomplete sentence alignments, duplication, language switching within sentences, repeated Arabic text, and formatting errors.

Collectively, these issues highlighted the need for optimizing alignment algorithms and incorporating manual interventions to ensure the accuracy and reliability of the generated sentence pairs. To address these challenges, manual interventions were employed to resolve cases where automated methods struggled, such as identifying, correcting, or removing misaligned pairs caused by cultural or linguistic variations. Additionally, flagged outputs with formatting errors or alignment issues were meticulously reviewed and corrected to ensure consistency and quality. These combined efforts not only improved the consistency and accuracy of alignments but also ensured that the generated pairs adhered more closely to contextual and semantic equivalence. The success of these interventions was validated by the high MQM quality scores obtained during human evaluations.

To strengthen the justification for using GPT, we conducted a comparative analysis of the embedding models employed in this study—Mpnet, LaBSE, Distiluse, and MiniLM—against existing sentence alignment methodologies. In comparison, *Chimoto & Bassett (2022)* used fine-tuned LaBSE embeddings to align sentences in their dataset, achieving 53.3% alignment accuracy and over 85% precision at the same similarity threshold. While our study shows that pre-trained models like Mpnet and LaBSE perform well without fine-tuning, it is important to acknowledge the differences in datasets. Their study focused on Luhya-English, which introduces unique alignment challenges, whereas our dataset of English-Arabic pairs benefits from broader representation in pre-trained models. This comparison highlights the adaptability of embedding models in different contexts and demonstrates the potential of combining them with GPT for generating accurate parallel sentences in resource-rich domains.

CONCLUSIONS

This study presents a comprehensive framework for leveraging the SauLTC and LLMs to generate high-quality parallel English-Arabic sentence datasets, addressing a critical gap in TE. By integrating advanced AI tools, such as GPT, and embedding models, the proposed methodology enhances the efficiency and accuracy of parallel sentence generation. The results highlight the effectiveness of this approach, with LaBSE achieving the highest alignment consistency. These findings suggest that AI can play a valuable role in supporting translation pedagogy and practice.

The results demonstrated the framework's ability to generate 15,845 sentence pairs, with LaBSE exhibiting the highest consistency and aligning 95% of the sentence pairs at the 0.7 threshold. Human evaluation further validated the dataset's quality with an MQM score of 98%, emphasizing its contextual accuracy and reliability in addressing alignment challenges.

Future directions will focus on expanding the framework to encompass the entire SauLTC, incorporating additional annotations and metadata to enhance the dataset's applicability across diverse domains. Furthermore, exploring advanced alignment techniques such as retrieval-augmented generation (RAG) can reinforce the contextual relevance and semantic accuracy of parallel sentence generation. Since automatic alignments and human evaluations present some challenges due to variations in linguistic interpretation and contextual nuances, future improvements should focus on fine-tuning embedding models to better capture language-specific variations. Additionally, developing context-aware scoring systems that integrate machine learning classifiers with human annotations will enhance AI evaluation metrics, ensuring more accurate and reliable assessments. By continuing to refine and expand this work, translation educators and researchers can equip students with the necessary skills to navigate the evolving landscape of AI-driven translation services, ultimately enriching the global exchange of knowledge and cultural understanding.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This research is funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R135), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Princess Nourah bint Abdulrahman University: PNURSP2025R135.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Moneerh Aleedy conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Fatma Alshihri conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Souham Meshoul conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Maha Al-Harhi conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Salwa Alramlawi conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Badr Aldaihani conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Hadil Shaiba conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Eric Atwell conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:
The code and data are available in the [Supplemental Files](#).

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj-cs.2788#supplemental-information>.

REFERENCES

- Alasmri I, Kruger H. 2018.** Conjunctive markers in translation from English to Arabic: a corpus-based study. *Perspectives* **26**(5):767–788 DOI [10.1080/0907676X.2018.1425463](https://doi.org/10.1080/0907676X.2018.1425463).
- Al-Batineh M, Al Tenaijy M. 2024.** Adapting to technological change: an investigation of translator training and the translation market in the Arab world. *Heliyon* **10**(7):e28535 DOI [10.1016/j.heliyon.2024.e28535](https://doi.org/10.1016/j.heliyon.2024.e28535).
- Albir A. 2017.** *Researching translation competence by PACTE group*. Amsterdam: John Benjamins Publishing Company.
- Alfarizy G, Mandala R. 2022.** Verification of unanswerable questions in the question answering system using sentence-BERT and cosine similarity. In: *2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2022*. Piscataway: IEEE DOI [10.1109/ICAICTA56449.2022.9932903](https://doi.org/10.1109/ICAICTA56449.2022.9932903).
- Al-Harathi M, Al-Saif A. 2019.** The design of the SauLTC application for the English-Arabic learner translation corpus. In: *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*. Association for Computational Linguistics, 80–88.
- Al-Harathi M, Alsaif A, Al-Nafjan E, Alshihri F, Saleh M. 2024.** Saudi learner translation corpus: the design and compilation of an English-Arabic learner translation corpus. *PLOS ONE* **19**(10):1–22 DOI [10.1371/journal.pone.0303729](https://doi.org/10.1371/journal.pone.0303729).
- Andrey K, Kunilovskaya M. 2014.** Russian learner translator corpus. In: Sojka P, Horák A, Kopeček I, Pala K, eds. *Text, Speech and Dialogue*. Cham: Springer International Publishing, 315–323.
- Bisiada M. 2017.** Universals of editing and translation. In: *Empirical Modelling of Translation and Interpreting*. Berlin: Language Science Press, 241–275.
- Bowker L. 2023.** *De-mystifying translation introducing translation to non-translators*. Oxfordshire: Routledge.
- Castagnoli S. 2020.** Translation choices compared: investigating variation in a learner translation corpus. In: Granger S, Lefer M-A, eds. *Translating and Comparing Languages: Corpus-Based Insights. Corpora and Language in Use Proceedings 6*. Louvain-la-Neuve: Presses universitaires de Louvain, 25–44.
- Castagnoli S, Ciobanu D, Kunz K, Kübler N, Volanschi A. 2011.** Designing a learner translator corpus for training purposes. In: Kübler N, ed. *Corpora, Language, Teaching, and Resources: From Theory to Practice*. Bern: Peter Lang, 221–248.
- Chimoto EA, Bassett BA. 2022.** Very low resource sentence alignment: Luhya and Swahili. In: *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*. Gyeongju, Republic of Korea: Association for Computational Linguistics, 1–8. Available at <https://aclanthology.org/2022.loresmt-1.1.pdf>.
- Church KW, Dagan I, Gale WA, Fung P, Helfman JI, Satish B. 1993.** Aligning parallel texts: do methods developed for English-French generalize to Asian languages? In: *Proceedings of the Pacific Asia Conference on Formal and Computational Linguistics, Taipei*, 112. Available at <https://api.semanticscholar.org/CorpusID:60082349>.
- Espunya A. 2014.** The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation* **48**:33–43 DOI [10.1007/s10579-013-9260-1](https://doi.org/10.1007/s10579-013-9260-1).
- Feng F, Yang Y, Cer D, Arivazhagan N, Wang W. 2020.** Language-agnostic BERT Sentence Embedding. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* **1**:878–891 DOI [10.18653/v1/2022.acl-long.62](https://doi.org/10.18653/v1/2022.acl-long.62).

- Freitag M, Foster G, Grangier D, Ratnakar V, Tan Q, Macherey W. 2021.** Experts, errors, and context: a large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics* **9**(1):1460–1474 DOI [10.1162/tacl_a_00437](https://doi.org/10.1162/tacl_a_00437).
- Garbovskiy N, Kostikova O. 2020.** Translation didactics: what are the ways to train a translation teacher? In: Zhao J, Li D, Tian L, eds. *Translation Education. New Frontiers in Translation Studies*. Singapore: Springer, 45–61 DOI [10.1007/978-981-15-7390-3_4](https://doi.org/10.1007/978-981-15-7390-3_4).
- Granger S, Lefer M-A. 2020a.** The multilingual student translation corpus: a resource for translation teaching and research. *Language Resources and Evaluation* **54**(4):1183–1199 DOI [10.1007/s10579-020-09485-6](https://doi.org/10.1007/s10579-020-09485-6).
- Granger S, Lefer M-A. 2020b.** Introduction: a two-pronged approach to corpus-based crosslinguistic studies. *Languages in Contrast* **20**:167–183 DOI [10.1075/lic.00014.int](https://doi.org/10.1075/lic.00014.int).
- Granger S, Lefer M-A. 2021.** Corpus-based translation studies: can we do better? Insights from a combined quantitative and qualitative survey. In: *UCCTS2021, Bertinoro (Italy), Book of Abstracts*.
- Grégoire F, Langlais P. 2018.** Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In: *COLING, 2018—27th International Conference on Computational Linguistics, Proceedings*, 1442–1453.
- Gwet KL. 2008.** Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology* **61**:29–48 DOI [10.1348/000711006X126600](https://doi.org/10.1348/000711006X126600).
- He Y. 2021.** Challenges and countermeasures of translation teaching in the era of artificial intelligence. *The 2nd International Conference on Computing and Data Science. Journal of Physics: Conference Series* **1881**(2021):022086 DOI [10.1088/1742-6596/1881/2/022086](https://doi.org/10.1088/1742-6596/1881/2/022086).
- Igualada M, Echeverri Á. 2019.** Because something should change. Translator and interpreter training: past, present and future. *MonTI* **11**:29–46 DOI [10.6035/MonTI.2019.11.1](https://doi.org/10.6035/MonTI.2019.11.1).
- Kanglang L, Afzaal M. 2021.** Artificial intelligence (AI) and translation teaching: a critical perspective on the transformation of education. *International Journal of Educational Sciences* **33**(1–3):64–73 DOI [10.31901/24566322.2021/33.1-3.1159](https://doi.org/10.31901/24566322.2021/33.1-3.1159).
- Kenny D. 2019.** Technology and translator training. In: *The Routledge Handbook of Translation and Technology*. Oxfordshire: Routledge, 498–515.
- Kong L. 2022.** Artificial intelligence-based translation technology in translation teaching. *Computational Intelligence and Neuroscience* **2022**:6016752 DOI [10.1155/2022/6016752](https://doi.org/10.1155/2022/6016752).
- Kruger H. 2012.** A corpus-based study of the mediation effect in translated and edited language. *Target. International Journal of Translation Studies* **24**:355–388 DOI [10.1075/target](https://doi.org/10.1075/target).
- Kübler N, Mestivier A, Pecman M. 2022.** Using comparable corpora for translating and post-editing complex noun phrases in specialized texts. In: Granger S, Lefer M-A, eds. *Extending the Scope of Corpus-Based translation Studies. Bloomsbury Advances in Translation*. London: Bloomsbury Publishing, 237–266.
- Kurek J, Latkowski T, Bukowski M, Świdorski B, Łepicki M, Baranik G, Nowak B, Zakowicz R, Dobrakowski Ł. 2024.** Zero-shot recommendation AI models for efficient job-candidate matching in recruitment process. *Applied Sciences* **14**(6):2601 DOI [10.3390/app14062601](https://doi.org/10.3390/app14062601).
- Lefer M-A. 2020.** Parallel corpora. In: Magali P, STh G, eds. *A Practical Handbook of Corpus Linguistics*. Cham: Springer International Publishing, 257–282.
- Lefer M-A, Piette J, Bodart R. 2022.** Machine translation post-editing annotation system (MTPEAS) manual. Available at <http://hdl.handle.net/20.500.12279/829>.

- Liu D.** 2022. IoT-based English translation teaching from the perspective of artificial intelligence. *International Journal of Antennas and Propagation* 2022:1749728 DOI 10.1155/2022/1749728.
- Lommel A, Gladkoff S, Melby A, Wright SE, Strandvik I, Gasova K, Vaasa A, Benzo A, Sparano RM, Foresi M, Innis J, Han L, Nenadic G.** 2024. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. ArXiv preprint DOI 10.48550/arXiv.2405.16969.
- Malcolm W.** 2004. *Translation quality assessment: an argumentation-centred approach*. Ottawa: University of Ottawa Press.
- Mohamed YA, Khanan A, Bashir M, Mohamed AHM, Adiel MAE, Elsadig MA.** 2024. The impact of artificial intelligence on language translation: a review. *IEEE Access* 12(3):25553–25579 DOI 10.1109/ACCESS.2024.3366802.
- Mohsen MA.** 2024. Artificial intelligence in academic translation: a comparative study of large language models and google translate. *Psycholinguistics* 35(2):134–156 DOI 10.31470/2309-1797-2024-35-2-134-156.
- O'Brien S.** 2012. Towards a dynamic quality evaluation model for translation. *Journal of Specialised Translation* 17(A):55–77.
- Park D, Padó S.** 2024. Multi-dimensional machine translation evaluation: model evaluation and resource for Korean. In: *LREC-COLING 2024*, 11723–11744.
- Reimers N, Gurevych I.** 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In: *EMNLP, 2020—2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 4512–4525 DOI 10.18653/v1/2020.emnlp-main.365.
- Rodríguez De Céspedes B.** 2019. Translator education at a crossroads: the impact of automation. *Lebende Sprachen* 64(1):103–121 DOI 10.1515/les-2019-0005.
- Saldías B, Foster G, Freitag M, Tan Q.** 2022. Toward more effective human evaluation for machine translation. In: *HumEval 2022—2nd Workshop on Human Evaluation of NLP Systems, Proceedings of the Workshop*, 76–89 DOI 10.18653/V1/2022.HUMEval-1.7.
- Sawyer D, Austerlühl F, Raído V.** 2019. *The evolving curriculum in interpreter and translator education: stakeholder perspectives and voices*. Amsterdam: John Benjamins Publishing Company.
- Song J, Zan H, Liu T, Zhang K, Ji X, Cui T.** 2024. Text classification based on multilingual back-translation and model ensemble. In: *Communications in Computer and Information Science 2080 CCIS*, 231–241 DOI 10.1007/978-981-97-1717-0_21.
- Steigerwald E, Ramírez-Castañeda V, Brandt DYC, Báldi A, Shapiro JT, Bowker L, Tarvin RD.** 2022. Overcoming language barriers in academia: machine translation tools and a vision for a multilingual future. *Bioscience* 72(10):988–998 DOI 10.1093/biosci/biac062.
- Uzar R, Waliński JT.** 2001. Analysing the fluency of translators. *International Journal of Corpus Linguistics* 6:155–166 DOI 10.1075/ijcl.
- Varga D, Halácsy P, Kornai A, Nagy V, Németh L, Trón V.** 2007. Parallel corpora for medium density languages. In: *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*. Amsterdam: John Benjamins Publishing Company, 247–258. Available at <https://benjamins.com/catalog/cilt.292>.
- Wang Y.** 2023. Artificial intelligence technologies in college English translation teaching. *Journal of Psycholinguistic Research* 52(5):1525–1544 DOI 10.1007/s10936-023-09960-5.
- Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL.** 2013. A comparison of Cohen's Kappa and Gwet's ACI when calculating inter-rater reliability coefficients: a study conducted with

personality disorder samples. *BMC Medical Research Methodology* **13(1)**:1–7
[DOI 10.1186/1471-2288-13-61](https://doi.org/10.1186/1471-2288-13-61).

Wu T, He S, Liu J, Sun S, Liu K, Han QL, Tang Y. 2023. A brief overview of ChatGPT: the history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* **10(5)**:1122–1136 [DOI 10.1109/JAS.2023.123618](https://doi.org/10.1109/JAS.2023.123618).

Wurm A. 2020. Translation quality in an error-annotated translation learner corpus. In: Granger S, Marie-Aude L, eds. *Translating and Comparing Languages: Corpus-based Insights*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Zaghlool ZDM, Khasawneh MAS. 2024. Aligning translation curricula with technological advancements; insights from artificial intelligence researchers and language educators. *Studies in Media and Communication* **12(1)**:58–70 [DOI 10.11114/smc.v12i1.6378](https://doi.org/10.11114/smc.v12i1.6378).

Zapata J. 2016. Investigating translator-information interaction: a case study on the use of the prototype biconcordancer tool integrated in CSMACAT. In: Carl M, Bangalore S, Schaeffer M, eds. *New Directions in Empirical Translation Process Research Exploring the CRITT TPR-DB*. Cham: Springer, 135–152 [DOI 10.1007/978-3-319-20358-4](https://doi.org/10.1007/978-3-319-20358-4).

Zhao J, Li D, Tian L. 2020. *Translation education*. Singapore: Springer.

Zhu S, Yang Y, Xu C. 2020. Extracting parallel sentences from nonparallel corpora using parallel hierarchical attention network. *Computational Intelligence and Neuroscience* **2020**:8823906
[DOI 10.1155/2020/8823906](https://doi.org/10.1155/2020/8823906).