p-ISSN 2541-0326

Volume: 11
Issue: 1
Year: September 2026
Page: 1-18

SCOPE
JOURNAL OF ENGLISH LANGUAGE TEACHING
Website: https://journal.lppmunindra.ac.id/index.php/SCOPE/
DOI: https://doi.org/10.30998/jh.v8i2.2784

Alignment and Bilingual Corpora in IT-Based Translation for English Language Teaching: A Systematic Literature Review

Refika Andriani[1], Fatma Setya Ningrum[2], Meutya Rizki Ramadhani[3], Aura Syalshabillah[4]

[1,2,3,4] English Education Department, Universitas Lancang Kuning, Pekanbaru, Indonesia
[1]refika@unilak.ac.id, [2]fatma.sn@unilak.ac.id, [3]meutyarizki@unilak.ac.id, [4]aurasyalsha15@gmail.com

Abstract. The increasing integration of information technology into translation studies has repositioned alignment and bilingual corpora as central resources for both linguistic analysis and English language teaching. Although extensive research has explored corpus-based translation and natural language processing, existing studies remain scattered across technical, pedagogical, and sociocultural perspectives. This systematic literature review addresses this fragmentation by synthesizing empirical evidence on alignment practices and bilingual corpus construction in IT-based translation. Guided by PRISMA procedures, twenty-two peer-reviewed journal articles published between 2020 and 2025 were retrieved from Scopus, Crossref, and Google Scholar and analyzed through thematic synthesis. The findings reveal a clear shift from manual and rule-based alignment toward semi-automated and hybrid models that integrate NLP techniques with human linguistic validation. While technological innovation has enhanced scalability, accuracy, and analytical efficiency, persistent challenges remain in domain adaptation, low-resource languages, and the interpretation of culturally embedded meaning. The review further demonstrates that bilingual corpora play a significant pedagogical role by supporting data-driven learning, learner autonomy, and critical language awareness in translation and English language teaching contexts. Beyond technical application, aligned corpora also function as sociocultural artifacts that reflect ideology, identity, and communicative purpose across languages. This study contributes an integrative framework that connects computational alignment, linguistic interpretation, and pedagogical application within a unified perspective. The findings highlight the importance of hybrid human–machine approaches for sustainable corpus development and suggest broader implications for translation education, ELT practice, and inclusive multilingual research in digitally mediated learning environments.

Keyword: Bilingual Corpus; Information Technology-Based Translation; Natural Language Processing

Received: 15-03-2026
Revised: 04-05-2026
Accepted: 30-09-2026
Corresponding Author: Refika Andriani; refika@unilak.ac.id

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

INTRODUCTION

In an era where language is increasingly mediated by technology, translation studies has emerged as a dynamic field that integrates linguistic inquiry with computational innovation. The rapid advancement of computational linguistics and information technology has not only transformed translation practices but also redefined how translation is conceptualized and taught. Rather than functioning solely as text producers, translators are now positioned as mediators of knowledge systems within digitally mediated environments. This transformation is particularly evident in the integration of neural machine translation and large language models, which reshape the relationship between human cognition and machine-assisted language processing (Alissa & Wald, 2023).

At the center of this transformation lies the concept of alignment, defined as the process of linking semantically equivalent segments across languages. Alignment plays a critical role in the construction of translation memories, bilingual corpora, and automated linguistic resources. It enables the systematic organization of multilingual data and supports the development of computational models that enhance translation quality. As digital translation technologies evolve, alignment has shifted from a purely technical procedure to a process that also reflects linguistic interpretation and meaning negotiation across languages.

Existing studies have demonstrated the importance of alignment and bilingual corpora in improving computational translation. Zhang (2022) highlights that alignment functions as a structural bridge between languages, contributing to coherence, accuracy, and consistency. Similarly, Bawden et al. (2021) emphasize that parallel corpora allow machine learning systems to learn translation correspondences with high precision, particularly for morphologically complex and low-resource languages. These developments indicate that bilingual corpora are not merely datasets but foundational infrastructures for both linguistic analysis and machine translation systems.

Recent research has further explored alignment practices across diverse language pairs, including English–Indonesian (Ayu Setiawati et al., 2021), Chinese–English (Yao & Zhang, 2023), and Arabic–English (Rahman & Al-Khateeb, 2024). Collectively, these studies demonstrate the expanding scope of corpus-based translation and its relevance across linguistic and cultural contexts. However, despite the growing body of research, the literature remains fragmented. Many studies are conducted within narrowly defined domains, focusing either on technical algorithm development or on linguistic analysis, without sufficiently integrating these perspectives. For example, Zhang (2022) examines structural alignment, while Alissa and Wald (2023) focus on transformer-based simplification, and Ar Rusyfa et al. (2025) address sociolinguistic dimensions of translation. This separation limits the development of a coherent and interdisciplinary understanding of alignment and corpus construction.

Furthermore, previous research has not adequately examined how the shift from rule-based systems to neural network architectures influences the construction and use of bilingual corpora. This gap is significant, as technological developments are not merely procedural changes but epistemological shifts that reshape how meaning, equivalence, and context are understood in translation. The absence of integrative syntheses that connect computational techniques, linguistic theory, and pedagogical applications highlights the need for a more comprehensive analytical framework.

Addressing this gap, the present study adopts an integrative approach that brings together computational modeling, corpus linguistics, and translation pedagogy. Unlike previous studies that

tend to separate technical and linguistic dimensions, this research conceptualizes alignment as both a linguistic and cognitive process that mediates between human interpretation and machine processing. This perspective represents the novelty of the study, positioning alignment as a multidimensional practice that operates across technological, linguistic, and educational domains.

The significance of this study lies in its contribution to both theoretical and practical aspects of translation studies. By synthesizing 22 peer-reviewed studies published between 2020 and 2025, this research aims to identify dominant alignment methods, commonly used technologies, and their pedagogical implications. The findings are expected to inform the development of translation tools, enhance teaching strategies, and provide a clearer understanding of how corpus-based approaches can support language learning and translation training.

The structure of this paper is organized as follows. The next section presents the methodological framework, including data sources, selection criteria, and analytical procedures used in this systematic literature review. This is followed by the results and discussion section, which synthesizes key findings related to alignment practices, technological developments, and pedagogical implications.

This study underscores the importance of integrating computational and linguistic perspectives in understanding alignment and bilingual corpus construction. By highlighting current trends, identifying research gaps, and proposing an integrative framework, this research contributes to advancing translation studies in digitally mediated contexts. It also offers broader implications for multilingual communication, computational linguistics, and English language teaching, particularly in promoting more inclusive and data-driven approaches to language learning and translation practice. Specifically, this study addresses three central questions:

1. What are the dominant alignment methods and corpus-building approaches in IT-based translation research (2020–2025)?
2. What technologies and software tools are most frequently applied?
3. What pedagogical and linguistic implications emerge from these practices?

METHOD

1. Data Sources and Search Strategy

The literature review was based on three major academic databases: Scopus, Crossref, and Google Scholar. These databases are interdisciplinary and include journals in linguistics, computational linguistics, and translation technology, thereby representing the most comprehensive sources of peer-reviewed literature on building bilingual corpora, aligning them, and translation through IT.

A Boolean search strategy was developed to balance breadth and precision in the literature search. Search terms included alignment, bilingual corpora, parallel corpora, IT-based translation, machine translation, and corpus construction. These terms were systematically combined using the Boolean operators AND and OR to focus the search on relevant literature on corpus alignment in translation and computational contexts.

To preserve the relevance and quality of the search, only English-language academic journal articles published between 2020 and 2025 were included. Any duplicate entries in the databases were removed, and the articles were then uploaded to a reference management software for subsequent screening and eligibility assessment.

2. Study Selection

The search used multiple scholarly databases, including Crossref, Google Scholar, Scopus, and Web of Science, to capture recent advances in research on bilingual corpora and translation automation. All selected documents met the inclusion criteria for systematic reviews, namely, peer-reviewed scholarly literature. By implementing a peer validation process, bias in study selection was mitigated, thereby enhancing the trustworthiness of the studies included.

For the search, the keywords and their various synonyms are listed in Table 2.1. These keywords were selected for their relevance to the study and their conceptual proximity to the research problem of corpus alignment, computational linguistics, and translation technology.

Table 2.1 Keywords of the Search Process

Keywords	Synonyms / Related Terms
Bilingual Corpus	parallel corpus, dual-language corpus, translation corpus, multilingual dataset
Alignment	corpus alignment, sentence alignment, text matching, bilingual mapping
IT-Based Translation	computer-assisted translation, machine translation, neural machine translation (NMT), computer-aided translation (CAT)
Corpus Linguistics	computational linguistics, data-driven linguistics, corpus-based analysis, text-mining linguistics
Natural Language Processing (NLP)	computational language processing, language technology, text mining, linguistic AI

3. Inclusion and Exclusion Criteria

To maintain the quality, relevance, and methodological rigor of this systematic review, specific inclusion and exclusion criteria were established prior to data screening. These criteria ensured that only studies directly addressing the review's objectives were selected. Each article was carefully evaluated based on its focus, methodological approach, and contribution to the field of bilingual corpus alignment and IT-based translation research. The criteria also helped to exclude studies that lacked empirical depth, methodological clarity, or scholarly reliability.

Table 2.2. Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Focused on bilingual or parallel corpus creation and alignment;	Addressed monolingual corpora;
Applied IT-based or computational tools for translation research; and	Described translation software without methodological testing; or
Provided empirical data or conceptual analysis.	Were non-peer-reviewed publications (e.g., theses, blogs, conference abstracts).

4. Screening and Selection Process

Initially, Crossref, Google Scholar, and Scopus brought in 1,700 records. After removing 270 duplicates and records deemed irrelevant or inaccessible, we continued the preliminary screening of 1,430 studies. Of the records screened, 1,197 were eligible for full-text review and were then evaluated for eligibility in accordance with PRISMA guidelines. Under both the inclusion and exclusion criteria, 233 papers consisting of book chapters and review articles were excluded, alongside 175 others that were irrelevant to the study focus and accessible records. After all revisions, 22 studies were selected for a thorough analysis and synthesis.

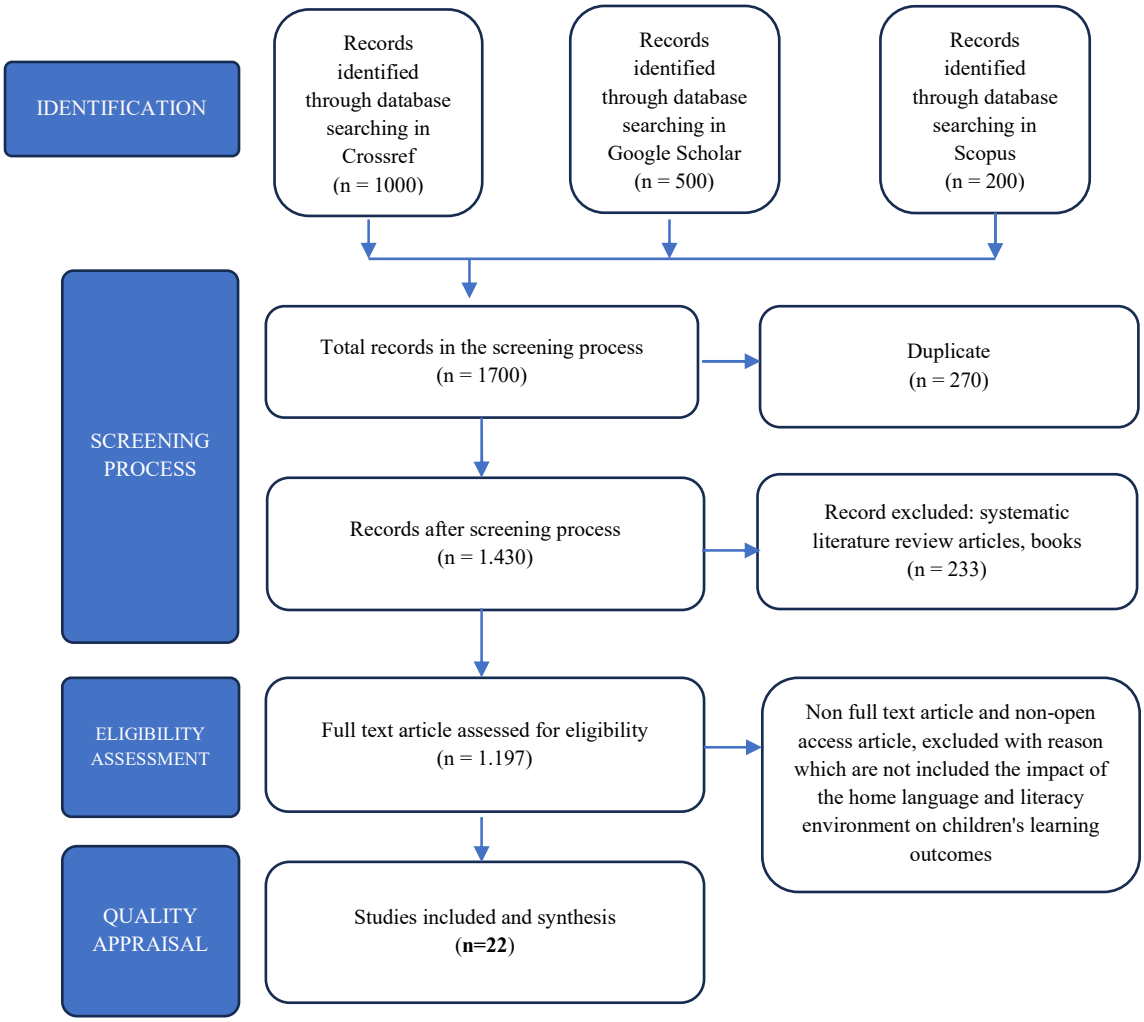


Figure 2.1. Article's selection procedures flowchart

The PRISMA-based screening process demonstrated a careful, systematic approach to identifying relevant studies. The first database search yielded a wide range of potential sources, which were then narrowed using a series of filters and a quality appraisal. The finalized dataset also benefited from the removal of duplicate and irrelevant records, thereby enhancing its reliability and validity. The final 22 studies form the empirical base of this review and were the most thematically and methodologically consistent research on aligning and constructing a bilingual corpus in IT-based translation.

5. Data Extraction and Analysis

To ensure consistency and transparency, data from the 22 included studies were extracted using a standardized coding form covering:

(1) research aim and focus
(2) type of corpus and language pairs
(3) level of alignment
(4) methods and technologies used
(5) research design, and
(6) key findings, including linguistic, computational, and pedagogical aspects.

A thematic synthesis approach was applied through four stages: coding, theme development, validation, and integration. First, all data were subjected to open coding, where each study was examined and segmented into meaningful units. Initial codes were assigned to capture key concepts such as alignment techniques, corpus construction, technological tools, and pedagogical applications.

Second, the codes were grouped into broader categories using an iterative comparison process. Similar codes were clustered into themes, including alignment practices, NLP-based technologies, and pedagogical implications, ensuring conceptual coherence and distinction between themes.

Third, theme validation was conducted by reviewing all themes across the dataset to ensure consistency and representativeness. Discrepancies were resolved through repeated examination of the data. To enhance reliability, the coding process was conducted iteratively and cross-checked for consistency.

Finally, the validated themes were linked to the research questions. Alignment-related themes addressed the first research question, technology-related themes addressed the second, and pedagogical themes addressed the third. This ensured that the analysis remained directly aligned with the study objectives. Overall, the thematic synthesis enabled a systematic interpretation of the literature by identifying key patterns, methodological trends, and pedagogical insights within IT-based translation studies.

5. Quality Assessment

To ensure methodological rigor and the reliability of the selected studies, this review employed the Critical Appraisal Skills Programme (CASP) framework. Given that the included studies consist of both qualitative and mixed-method research, this study primarily adopted the CASP Qualitative Checklist as the main appraisal tool, as it provides flexible criteria applicable to diverse research designs. The checklist was adapted to allow consistent evaluation across studies with varying methodological approaches in bilingual corpus alignment and IT-based translation research.

Each study was assessed using ten core CASP domains, including research clarity, methodological appropriateness, sampling adequacy, data collection procedures, analytical rigor, consideration of bias, validity of findings, relevance, and overall contribution to the field. These domains were operationalized into a 10-point scoring system to facilitate comparison across studies.

Based on the scoring criteria, studies were categorized as high quality (8–10), moderate quality (5–7), or low quality (below 5, excluded). Out of the 1,700 identified records, 22 studies met all inclusion criteria and achieved a minimum score of 5, with a mean quality score of 8.2. This indicates a generally high level of methodological quality among the selected studies.

To ensure consistency, the appraisal process was conducted systematically across all studies, with attention to variations in research design, corpus type, and analytical approaches. The use of the CASP framework helped reduce potential bias and provided a structured basis for evaluating the reliability and relevance of the included literature.

RESULTS AND DISCUSSION

This review organizes and synthesizes twenty-two empirical studies conducted between 2020 and 2025, focusing on alignment and the construction of bilingual corpora in IT-based translation. The studies span a variety of linguistic pairs, methods, and technologies, thereby confirming the complexity of translation studies. The general bibliographic information for the studies, including authors, article titles, journals, and years of publication, is presented in Table 3.1. This overview

serves as the foundation for answering the research questions and for discussing key patterns emerging from the literature.

Table 3.1. Summary of Reviewed Studies (2020–2025)

No	Study	Objective	Language Pair / Corpus	Alignment Method	Tools / Technology	Key Findings
1	**Alissa & Wald (2023)**. *Text Simplification Using Transformer and BERT*. CMC.	Improve text simplification using Transformer and BERT	English parallel corpus (WikiSmall)	Sentence-level alignment (NMT)	Transformer, BERT, SentencePiece	Transformer improves performance; BERT causes overfitting
2	**Ar Rusyfa et al. (2025)**. *Positioning in News Translation Regarding Hamas: A Parallel Corpus-Based Approach*. JOLLT	To examine how Hamas is positioned in translated news within the Israel–Palestine conflict using a parallel corpus approach	English–Indonesian parallel corpus (160 translated news articles from Reuters and Rep)	Corpus-based alignment and discourse analysis (CDA, Appraisal Theory, Narrative Theory)	Sketch Engine (corpus tool), corpus-based analysis techniques	Translators apply "redacted labelling" (omission, addition, softening) to shift Hamas representation from negative to more neutral, influenced by media ideology and socio-political context
3	**Ardi et al. (2025)**. *L1 Interference in Students' Translations: A Corpus-Based Analysis of Collocation Errors and Pedagogical Implications*. International Journal of Language Pedagogy	To analyze collocation errors in student translations and identify the influence of L1 interference in Indonesian–English translation	Indonesian–English learner corpus (24 student translations of Minangkabau cultural text) compared with reference corpora (BNC, COCA)	Corpus-based error analysis (collocation classification and comparison with reference corpora)	AntConc, Kortara, BNC, COCA, English Collocation Dictionary	151 collocation errors identified, dominated by noun–noun errors (58.94%); errors mainly caused by L1 interference and literal translation; highlights need for corpus-based pedagogy and explicit collocation instruction
4	**Ayu Setiawati et al. (2021)**.	To analyze lexical equivalence	English–Indonesian parallel	Sentence-level alignment and concordance-	AntPConc (v1.2.1), parallel	Found 23 verb equivalents and 8 noun

	The Translation's Equivalence of the Word "Support" in Parallel Corpus English–Indonesian. ACIELJ	and distribution of the word "support" in English–Indonesian parallel corpus and examine its translation at the lexical level	corpus (film subtitles; 4.8M tokens; aligned bilingual corpus)	based lexical analysis	corpus dataset (GitHub), concordance analysis	equivalents of "support"; translation is mostly consistent with lexical class (verb→verb, noun→noun); high frequency variation shows semantic flexibility and context-dependent equivalence
5	**Bawden et al. (2021)**. *DiaBLa: A Corpus of Bilingual Spontaneous Written Dialogues for Machine Translation*. Language Resources & Evaluation	To develop a bilingual dialogue corpus for evaluating machine translation quality and analyzing MT-mediated communication	English–French bilingual dialogue corpus (144 spontaneous dialogues; ~5,700 sentences; MT-mediated chat data)	Sentence-level alignment with dialogue context (context-aware alignment across turns)	Neural MT systems (baseline & context-aware 2-to-2 model), Nematus, Marian, OpenSubtitles training data	Provides a unique MT evaluation corpus with human sentence-level judgments; shows context-aware MT improves coherence; identifies key MT challenges such as word choice, coherence, and dialogue consistency
6	**Bullock, Serigos, & Toribio (2021)**. *Exploring a Loan Translation and Its Consequences in an Oral Bilingual Corpus*. Journal of Language Contact	To investigate whether the construction *agarrar* + *NP* in Spanish is a loan translation (calque) from English *get* + *NP* and to analyze its linguistic and grammatical consequences	Spanish–English bilingual corpus (Spanish in Texas Corpus, ~500,000 words) compared with monolingual corpora (Corpus del Español, activ-es subtitles, SubtlexUS)	Frequency-based comparative analysis + corpus comparison (cross-corpus alignment of lexical patterns and collocations)	Python scripts for corpus extraction, manual annotation, statistical analysis (Chi-square, logistic regression), corpus platforms	Finds that *agarrar* + *NP* is a contact-induced loan translation from English; frequency similar across corpora but usage differs qualitatively; in Texas Spanish it occurs mostly with abstract nouns (support verb construction); demonstrates

						that loan translation can reshape grammar, not just vocabulary
7	Chen (2023). *An Improved Corpus-Based NLP Method for Facilitating Keyword Extraction*. Sustainability.	Improve keyword extraction accuracy by integrating multi-variable NLP approach	Monolingual scientific corpus (200 SCI articles on COVID-19 vaccine hesitancy)	Corpus-based statistical alignment (log-likelihood + i10-index + IPA clustering)	AntConc, COCA corpus, NGSL, function word list	Quadrant I keywords (n=98) achieve highest lexical coverage (9.81%); method improves accuracy and reduces noise in keyword extraction
8	Chen, Chang, & Yang (2022). *Integrating corpus-based and NLP approach to extract terminology and domain-oriented information*. Acta Scientiarum Technology.	To integrate corpus-based and NLP approaches to extract terminology and domain knowledge from specialized corpus	Monolingual domain corpus (US Army FM 8-10-6 military text)	Corpus-based keyword alignment + rule-based NLP parsing (hybrid alignment approach)	AntConc 3.5.8, CRS (Computer Reading System), log-likelihood method	Integrated approach successfully identifies domain terminology, builds text database, and extracts domain knowledge; outperforms traditional corpus methods that cannot capture semantic meaning
9	Dang (2024). *The extraction method used for English–Chinese machine translation corpus based on bilingual sentence pair coverage*. Open Computer Science.	To improve bilingual corpus extraction for machine translation using sentence pair coverage and machine learning	English–Chinese bilingual parallel corpus	Sentence-level alignment using bilingual sentence pair coverage + machine learning filtering	Statistical machine translation (phrase-based, hierarchical, syntax-based MT), alignment modules, preprocessing and MERT optimization	Proposed method improves corpus quality by filtering noisy and misaligned sentence pairs, leading to better translation accuracy and performance in ECMT systems
10	Endarto (2020). *A*	To identify lexical	English corpus from	No alignment (monolingual	Corpus Pattern	Indonesian English is

	corpus-based lexical analysis of Indonesian English as a new variety. Indonesian Journal of Applied Linguistics.	characteristics of Indonesian English as a potential new variety using corpus-based analysis	Indonesian online newspaper (The Jakarta Post)	corpus analysis using lexical patterns)	Analysis (CPA), #LancsBox corpus tool	characterized by four lexical features: loanwords, semantic shifts, morphological shifts, and unique collocations influenced by L1 (Indonesian), indicating emergence of a localized English variety
11	Luo & Xiang (2020). *Application of Data Mining Methods in Internet of Things Technology for the Translation Systems in Traditional Ethnic Books*. IEEE Access.	To improve machine translation of ethnic texts using machine learning and IoT through a discontinuous phrase extraction model	Bilingual corpus (e.g., Uyghur–Chinese, minority language corpora)	Phrase-level alignment (statistical MT with discontinuous phrase extraction)	Log-linear model, Statistical Machine Translation (SMT), Machine Learning, IoT, BLEU & NIST evaluation	Discontinuous phrase extraction model outperforms traditional methods by extracting more phrases, improving translation fluency and achieving higher BLEU and NIST scores
12	Matvieieva (2020). *Translation of Legal Terms: Bilingual Dictionary vs Parallel Corpus*. Science and Education: A New Dimension: Philology.	To compare the effectiveness of bilingual dictionaries and parallel corpora in translating legal terms	Ukrainian–English parallel legal corpus	Sentence-level alignment in parallel corpus (context-based equivalence analysis)	Parallel corpus analysis, component analysis, bilingual dictionaries	Parallel corpora provide more accurate and context-sensitive translations than dictionaries; corpus-based translation improves semantic understanding, term selection, and translation quality
13	Pakaja, Pong, Som, & Sari (2025). *Development*	To develop a large, linguistically annotated	Indonesian–English parallel corpus (multi-	Sentence-level alignment (manual + semi-	NLP tools (NLTK, SpaCy), Sentence	Corpus achieves high alignment accuracy

	of an Indonesian-English Parallel Corpus for Translation and Comparative Linguistics Research. JILTECH.	Indonesian–English parallel corpus for translation and comparative linguistic analysis	genre: literary, news, academic, informal texts; ~1 million words, ~25,000 sentence pairs)	automatic) with syntactic and semantic annotation	Aligner, Transalign, machine learning algorithms	(92%) and supports analysis of translation equivalence, syntactic transformation, and word order differences; improves machine translation performance, especially for formal texts
14	Pakaja, Pong, Som, & Sari (2025). *Development of an Indonesian-English Parallel Corpus for Translation and Comparative Linguistics Research*. JILTECH.	To develop a large, linguistically annotated Indonesian–English parallel corpus for translation and comparative linguistic analysis	Indonesian–English parallel corpus (multi-genre: literary, news, academic, informal texts; ~1 million words, ~25,000 sentence pairs)	Sentence-level alignment (manual + semi-automatic) with syntactic and semantic annotation	NLP tools (NLTK, SpaCy), Sentence Aligner, Transalign, machine learning algorithms	Corpus achieves high alignment accuracy (92%) and supports analysis of translation equivalence, syntactic transformation, and word order differences; improves machine translation performance, especially for formal texts
15	Quick & Hartmann (2021). *The Building Blocks of Child Bilingual Code-Mixing: A Cross-Corpus Traceback Approach*. Frontiers in Psychology.	To investigate how bilingual children construct code-mixed utterances based on input patterns and individual differences	German–English bilingual child corpus (longitudinal data of two children aged 2–4; monolingual + code-mixed utterances)	Pattern-based alignment using traceback method (within-corpus and cross-corpus comparison)	Traceback algorithm, corpus-based analysis, bigram network analysis	Around 50% of code-mixed utterances can be traced to input patterns; code-mixing is structured around frame-and-slot constructions; strong influence of caregiver input and individual differences
16	Rajeg & Rajeg (2022). *A Corpus*	To measure constructional equivalence	English–Indonesian parallel	Sentence-level alignment combined with	R programming (paracorp),	High degree of constructional equivalence;

	Linguistic Study of Constructional Equivalence for the Indonesian Translation of ROB and STEAL. Parole Journal.	of near-synonyms (ROB vs STEAL) in English-Indonesian translation using corpus-based approach	corpus (OpenSubtitles; 9.7 million aligned sentences)	constructional alignment (participant roles & grammatical constructions)	parallel concordance (KWIC), statistical analysis (Chi-square, Fisher-Yates)	ROB profiles TARGET while STEAL profiles GOODS; Indonesian translations largely preserve these patterns, though minor shifts occur
17	Saddhono et al. (2023). *Corpus Linguistics Use in Vocabulary Teaching Principle and Technique Application: A Study of Indonesian Language for Foreign Speakers*. IJSCL.	To analyze vocabulary, terminology, and grammar of BIPA students and develop effective vocabulary teaching principles using corpus linguistics	Indonesian learner corpus (BIPA students' written texts; 51 students from Fatoni University & Yale University)	No explicit bilingual alignment; monolingual corpus analysis (frequency & concordance-based analysis)	Kortara (Korpus Nusantara), corpus frequency analysis, concordance (KWIC), mixed-method approach	Vocabulary mastery dominated by nouns; 1,947 word types and 9,664 tokens; 11 high-frequency words identified; corpus analysis supports effective vocabulary teaching principles and techniques for BIPA learners
18	Shawaqfeh & Khasawneh (2023). *Incorporating Corpus Linguistics Tools in the Training and Professional Development of Lecturers in Translation Studies*. Studies in Media and Communication.	To evaluate the impact and strategies of integrating corpus linguistics tools in the professional development of translation lecturers	No specific corpus dataset; based on survey data from 114 translation lecturers and PhD students (multi-language context)	No alignment (focus on pedagogical integration, not corpus alignment)	AntConc, TransTool, Sketch Engine, WordSmith Tools, web-based corpora, corpus-based tasks	Corpus tools significantly enhance translation teaching; AntConc widely used (92.1%) and highly impactful; TransTool supports domain-specific translation; corpus-based tasks improve pedagogical effectiveness; mixed perceptions on web-based corpora

19	Sianipar & Sajarwa (2022). The Lexical Equivalence Variety of F-Words Translation into Indonesian: Corpus-Based Translation Research. Lire Journal.	To identify lexical equivalence variations and functional meanings of F-Words translation in Indonesian subtitles.	English–Indonesian parallel corpus (OpenSubtitles 2016; ~4.3 million sentence pairs).	Sentence-level alignment (parallel subtitle corpus).	AntConc software (v5.3.8), corpus-based analysis, qualitative–quantitative approach.	F-Words show high lexical variation (7938 occurrences); translations are not always taboo; functions include emotion, emphasis, solidarity, insult, failure, and misery; meaning depends on context and collocation.
20	Vîlceanu (2023). *Integrating Interview-Based Approaches into Corpus-Based Translation Studies and Literary Translation Studies*. Open Linguistics.	To develop an integrative framework combining corpus-based translation studies and literary translation studies through interview-based corpora (paratexts)	Interview-based corpus (mediated interviews with literary translator; Romanian context; 2009–2023 dataset)	No traditional alignment; qualitative corpus design using paratextual data (interview corpus as complementary to textual analysis)	Qualitative methods (mediated interviews, discourse analysis), corpus design framework, digital humanities perspective	Interview corpora provide rich data on translator identity, cognition, and decision-making; paratexts complement corpus analysis; integration of qualitative and corpus-based methods enhances understanding of translation processes and bridges theory–practice gap
21	Xu (2023). *Translation Studies from the Perspective of Corpus Translation Studies and Digital Humanities*. Journal of Education,	To explore the role of corpus translation studies and digital humanities in shaping translation theory, methodology, and research paradigms	Not specific (conceptual discussion of bilingual and multilingual corpora in translation studies)	Conceptual discussion of alignment (sentence-level correspondence in parallel corpora; no empirical alignment implementation)	Digital humanities technologies (big data, AI, text mining, machine learning), parallel corpus, translation memory	Corpus translation studies promote data-driven translation research; alignment and parallel corpora are central to improving

	Teaching and Social Studies.					translation quality and efficiency; digital humanities enable paradigm shifts, theoretical innovation, and interdisciplinary integration in translation studies
22	Zhang (2022). *Design of New Word Retrieval Algorithm for Chinese-English Bilingual Parallel Corpus*. Mathematical Problems in Engineering.	To design a sentence alignment system and new word retrieval approach for improving Chinese–English bilingual corpus construction and machine translation	Chinese–English bilingual corpus (web-based data from CNKI and bilingual web pages)	Sentence-level alignment using length-based algorithm combined with evaluation function and dynamic programming	Length-based alignment model, StarDict dictionary, stemming technique, evaluation metrics (precision, recall, F-score)	The proposed alignment system improves precision and recall; effectively handles unknown words; supports cross-language information retrieval and machine translation, though corpus imbalance remains a challenge

As shown in Table 3.1, the reviewed studies reveal substantial variation in research scope and disciplinary focus. The distribution across international journals such as *Computational Linguistics*, *Languages*, *Open Linguistics*, and *International Journal of Society, Culture and Language* demonstrates that bilingual corpus research has become increasingly global and interdisciplinary. The diversity of data sources and analytical approaches also reflects a balance between linguistic inquiry, computational modeling, and pedagogical application. Based on this mapping, the results and discussion are organized according to the three research questions of this study.

1. **Alignment Practices and Corpus Construction Approaches**
   The reviewed studies indicate a clear shift in alignment practices from manual and rule-based procedures toward semi-automated and hybrid models. Research on English–Indonesian, Chinese–English, and Arabic–English corpora demonstrates that contemporary alignment increasingly relies on multi-level strategies, with sentence-level alignment emerging as the most dominant approach across studies. This is often complemented by word-level and phrase-level alignment to capture finer linguistic units. In more advanced models, semantic alignment and hybrid alignment approaches are applied to better represent contextual meaning

and cross-linguistic equivalence (Ayu Setiawati et al., 2021; Rajeg & Rajeg, 2022; Zhang, 2022; Dang, 2024). These approaches enable researchers to process large-scale bilingual data while preserving sensitivity to meaning equivalence and contextual nuance, which are essential for both research accuracy and pedagogical relevance.

In parallel, corpus construction is no longer viewed as a purely technical process. Several studies emphasize that corpus design involves linguistic and methodological decisions related to genre selection, annotation schemes, and validation procedures (Matvieieva, 2020; Bawden et al., 2021; Pakaja et al., 2025). These decisions directly affect the interpretability and usability of corpora in translation studies and English language teaching. As a result, alignment emerges as a central analytical practice that mediates between computational efficiency and human interpretation.

2. Technological Tools and NLP-Based Innovation

Technological innovation constitutes a dominant theme across the reviewed literature, particularly through the integration of natural language processing techniques. Many studies employ NLP-based tools for sentence alignment, terminology extraction, keyword identification, and semantic similarity measurement, which significantly enhance scalability and analytical precision (Chen, 2023; Chen et al., 2022; Zhang, 2022). Most NLP-based studies primarily operate at the sentence level, with some extending to word-level and semantic alignment through similarity-based techniques. These tools allow bilingual corpora to function as dynamic analytical resources rather than static data repositories.

Despite these advances, the findings also reveal persistent limitations associated with technology-driven alignment. Several studies report challenges related to domain specificity, idiomatic language, and culturally embedded meanings that automated systems struggle to capture (Giampieri, 2023; Kwok et al., 2023; Leńko-Szymańska & Biel, 2023). Research conducted in lower-resource contexts further highlights the need for human validation to ensure reliability and interpretive depth (Endarto, 2020; Sianipar & Sajarwa, 2022). Consequently, the dominant technological trend favors hybrid workflows that combine automated processing with linguistic expertise.

3. Pedagogical and Linguistic Implications for Translation and English Language Teaching

A significant number of reviewed studies emphasize the pedagogical value of bilingual corpora in translation education and English language teaching. Corpus-based approaches support data-driven learning by enabling learners to work with authentic language data, analyze collocation patterns, and justify translation choices through empirical evidence (Ardi et al., 2025; Saddhono et al., 2023; Shawaqfeh & Khasawneh, 2023). Such practices foster learner autonomy, critical thinking, and metalinguistic awareness.

Beyond technical skill development, several studies highlight the role of bilingual corpora in promoting intercultural understanding. Research on news translation, taboo language, and legal discourse illustrates how alignment reveals ideological positioning and sociocultural values embedded in texts (Ar Rusyfa et al., 2025; Sianipar & Sajarwa, 2022; Vîlceanu, 2023). These findings suggest that corpus-based translation can function as a bridge between linguistic analysis and cultural interpretation. For English language teaching contexts, this reinforces the importance of integrating corpus tools in ways that support both linguistic competence and critical literacy.

CONCLUSION

This systematic literature review demonstrates that alignment and bilingual corpus construction have evolved into core components of IT-based translation research, functioning at the intersection of computational technology, linguistic analysis, and educational practice.

Synthesizing twenty-two studies published between 2020 and 2025, the findings confirm that contemporary alignment practices are no longer limited to technical sentence matching but operate as multi-level, interpretive processes that shape how meaning is negotiated across languages. The reviewed literature consistently shows a shift toward semi-automated and hybrid alignment models that integrate natural language processing with human linguistic judgment, particularly when corpora are designed for pedagogical and applied purposes.

The results further indicate that technological innovation, especially through NLP-based tools, has significantly expanded the scale and analytical capacity of bilingual corpora. At the same time, these advancements remain constrained by domain specificity, linguistic diversity, and cultural nuance, reinforcing the need for human validation and contextual interpretation. This balance between automation and expertise emerges as a defining feature of sustainable corpus development in translation studies. Rather than replacing linguistic insight, technology functions most effectively as a support system that enhances analytical precision and pedagogical usability.

From an educational perspective, the review highlights that bilingual corpora play a vital role in translation training and English language teaching. Corpus-based approaches promote data-driven learning, learner autonomy, and critical language awareness by enabling learners to engage directly with authentic linguistic evidence. Moreover, the sociocultural dimension of aligned corpora underscores that translation is not a neutral activity but a culturally situated practice that reflects ideology, identity, and communicative purpose.

Overall, this study contributes an integrated perspective that positions bilingual corpora as both technological infrastructures and pedagogical instruments. By linking alignment practices, NLP innovation, and educational application within a single analytical framework, the review advances understanding of IT-based translation as a hybrid human–machine endeavor. Future research should prioritize inclusive corpus development for low-resource languages, longitudinal evaluation of pedagogical impact, and ethical considerations in corpus design, ensuring that technological progress aligns with linguistic diversity and educational equity.

REFERENCE

Ar Rusyfa, I. Z., Datang, F. A., & Junaidi, J. (2025). Positioning in news translation regarding Hamas: A parallel corpus-based approach. *JOLLT Journal of Languages and Language Teaching, 13*(1), 60–73. https://doi.org/10.33394/jollt.v13i1.13445

Ardi, H., Ermanto, N., Juita, N., & Vy Rany. (2025). L1 interference in students' translations: A corpus-based analysis of collocation errors and pedagogical implications. *International Journal of Language Pedagogy, 5*(1), 93–105. https://doi.org/10.24036/ijolp.v5i1.113

Ayu Setiawati, N. M., Rajeg, I. M., & Matradewi, N. K. W. (2021). The translation's equivalence of the word "support" in the parallel corpus English–Indonesian. *Alford Council of International English & Literature Journal (ACIELJ), 4*(4), 1–10. https://www.acielj.com

Bawden, R., Chaumartin, F.-R., Dymetman, M., Marie, B., & Wisniewski, G. (2021). DiaBLa: A corpus of bilingual spontaneous written dialogues for machine translation. *Computational Linguistics, 47*(4), 771–806. https://doi.org/10.1162/coli_a_00416

Bullock, B. E., Gerfen, C., & Toribio, A. J. (2021). Exploring a loan translation and its consequences in an oral bilingual corpus. *Languages, 6*(3), 137. https://doi.org/10.3390/languages6030137

Chen, L.-C. (2023). An improved corpus-based NLP method for facilitating keyword extraction: An example of the COVID-19 vaccine hesitancy corpus. *Sustainability, 15*(4), 3402. https://doi.org/10.3390/su15043402

Chen, L.-C., Chang, K.-H., & Yang, S.-C. (2022). Integrating corpus-based and NLP approaches to extract terminology and domain-oriented information: An example of the US military corpus. *Acta Scientiarum Technology, 44*(1), e60486. https://doi.org/10.4025/actascitechnol.v44i1.60486

Dang, P. (2024). The extraction method used for the English-Chinese machine translation corpus is based on bilingual sentence pair coverage. *Open Computer Science, 14*, 20230107. https://doi.org/10.1515/comp-2023-0107

Endarto, I. T. (2020). A corpus-based lexical analysis of Indonesian English as a new variety. *Indonesian Journal of Applied Linguistics, 10*(1), 95–106. https://doi.org/10.17509/ijal.v10i1.24993

Espinosa-Anke, L., Palmer, G., Corcoran, P., Filimonov, M., Spasić, & Knight, D. (2021). English–Welsh cross-lingual embeddings. Applied Sciences, 11(14), 6541. https://doi.org/10.3390/app11146541

Fonseca, C., Guelpeli, M., & Netto, R. (2022). Representation of structured data of the text genre as a technique for automatic text processing. Texto Livre Linguagem e Tecnologia, 15, e35445. https://doi.org/10.35699/1983-3652.2022.35445

Giampieri, P. (2023). Is machine translation reliable in the legal field? A corpus-based critical comparative analysis for teaching purposes. *Languages, 8*(3), 194. https://doi.org/10.3390/languages8030194

Gorozhanov, A., Guseynova, I., & Степанова, Д. (2024). Natural language processing and fiction text: basis for corpus research. Rudn Journal of Language Studies Semiotics and Semantics, 15(1), 195–210. https://doi.org/10.22363/2313-2299-2024-15-1-195-210

Guo, Z. & Jifeng, Z. (2021). Japanese translation teaching corpus based on a bilingual non parallel data model. Journal of Intelligent & Fuzzy Systems, 40(2), 3731-3741. https://doi.org/10.3233/jifs-189407

Hu, L. and Hu, J. (2022). An exploration of the problems and solutions involved in translating computer software into Japanese. Mathematical Problems in Engineering, 2022, 1-9. https://doi.org/10.1155/2022/3712090

in, J., Tan, T., & Malancon, B. (2024). Hybrid distance-statistical-based phrase alignment for analyzing parallel texts in standard malay and malay dialects. Malaysian Journal of Computer Science, 37(1), 1-25. https://doi.org/10.22452/mjcs.vol37no1.1

Inácio, M., Cabezudo, M., Ramisch, R., Felippo, A., & Pardo, T. (2023). The amr-pt corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. Delta Documentação De Estudos Em Lingüística Teórica E Aplicada, 39(3). https://doi.org/10.1590/1678-460x202339355159

Kwok, H. L., Laviosa, S., & Liu, K. (2023). Lexical simplification in learner translation: A corpus-based approach. *Research in Corpus Linguistics, 11*(2), 103–124. https://doi.org/10.32714/ricl.11.02.06

Leńko-Szymańska, A., & Biel, Ł. (2023). Terminological collocations in trainee and professional legal translations: A learner-corpus study of L2 company law translations. *International Journal of Learner Corpus Research, 9*(1), 29–60. https://doi.org/10.1075/ijlcr.22008.len

Li, Y. (2023). Key technologies for constructing bilingual corpus for english-chinese translation. The Frontiers of Society Science and Technology, 5(11). https://doi.org/10.25236/fsst.2023.051109

Li, Y. and Li, Z. (2023). Research on college english translation teaching based on parallel corpus. International Journal of New Developments in Education, 5(18). https://doi.org/10.25236/ijnde.2023.051818

Matvieieva, S. A. (2020). Translation of legal terms: Bilingual dictionary vs. parallel corpus. *Science and Education a New Dimension: Philology, 8*(67), 45–49. https://doi.org/10.31174/SEND-Ph2020-225VIII67-10

Pakaja, M., Pong, M., Som, R., & Sari, H. F. A. (2025). Development of an Indonesian–English parallel corpus for translation and comparative linguistics research. *Journal International of Lingua and Technology (JILTECH), 4*(1), 16–30. https://doi.org/10.55849/jiltech.v4i1.817

Quick, A. E., & Hartmann, S. (2021). The building blocks of child bilingual code-mixing: A cross-corpus traceback approach. *Frontiers in Psychology, 12*, 682838. https://doi.org/10.3389/fpsyg.2021.682838

Rajeg, I. M., & Rajeg, G. P. (2022). A corpus linguistic study of constructional equivalence for the Indonesian translation of ROB and STIMULUS-EXPERIENCER verbs in English. *Linguistics and Literature Studies, 10*(2), 33–48. https://doi.org/10.13189/lls.2022.100202

Saddhono, K., Rohmadi, M., Setiawan, B., Suhita, R., Rakhmawati, A., Hastuti, S., & Islahuddin, I. (2023). Corpus linguistics use in vocabulary teaching principle and technique application: A study of Indonesian language for foreign speakers. *International Journal of Society, Culture & Language, 11*(1), 232–245. https://doi.org/10.22034/ijscl.2022.1971972.2823

Shawaqfeh, A. T., & Khasawneh, M. A. S. (2023). Incorporating corpus linguistics tools in the training and professional development of lecturers in translation studies. *Studies in Media and Communication, 11*(7), 260–273. https://doi.org/10.11114/smc.v11i7.6379

Sianipar, I. F., & Sajarwa. (2022). The lexical equivalence variety of F-words translation into Indonesian: Corpus-based translation research. *Lire Journal (Journal of Linguistics and Literature), 6*(1), 53–70. https://doi.org/10.33019/lire.v6i1.140

Tan, T., Lim, C., & Rahman, W. (2021). Sliding window and parallel lstm with attention and cnn for sentence alignment on low-resource languages. Pertanika Journal of Science and Technology, 30(1), 97-121. https://doi.org/10.47836/pjst.30.1.06

Tian, C. (2023). The translation of culture-specific items (csis) in yellow river cultural corpus: a case study of history of the yellow river. English Language Teaching and Linguistics Studies, 5(3), p381. https://doi.org/10.22158/eltls.v5n3p381

Tran, P., Nguyen, T., Vu, D., Tran, H., & Vo, B. (2022). A method of chinese-vietnamese bilingual corpus construction for machine translation. Ieee Access, 10, 78928-78938. https://doi.org/10.1109/access.2022.3186978

Vanhaverbeke, M., Enghels, R., Couto, M., & Ivanova, I. (2025). Enhancing code-switching research through comparable corpora: introducing the el paso bilingual corpus. Languages, 10(7), 174. https://doi.org/10.3390/languages10070174

Vîlceanu, T. (2023). Integrating interview-based approaches into corpus-based translation studies and literary translation studies. *Open Linguistics, 9*, 20220265. https://doi.org/10.1515/opli-2022-0265

Xia, L. and Zhang, H. (2024). A study on the construction of a large-scale multimodal parallel corpus of typical official documents of China, 1(10), 72-76. https://doi.org/10.62381/p243a12

Xu, J. (2023). Translation studies from the perspective of corpus translation studies and digital humanities. *Journal of Education, Teaching and Social Studies*. https://doi.org/10.22158/jets.v5n4p45

Zhang, D. (2022). Research on process of the bilingual corpus alignment tool. Learning & Education, 10(5), 35. https://doi.org/10.18282/l-e.v10i5.2664

Zhang, L. (2022). Design of new word retrieval algorithm for Chinese–English bilingual parallel corpus. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2022/4593129