Int. j. adv. multidisc. res. stud. 2026; 6(2):1567-1573

Received: 27-02-2026
Accepted: 07-04-2026

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Empowering Translation Pedagogy in the New Liberal Arts Era: An AI-TPACK Framework for Course Design and Digital Resource Construction in Corpus Processing

[1] Zhong Yucheng, [2] Song Qijun

[1, 2] School of International Studies, Guangdong University of Education, Guangzhou, Guangdong Province, China

Corresponding Author: **Zhong Yucheng**

Abstract

Situated within the macro-policy context of China's "New Liberal Arts" mandate, this article conceptualizes and evaluates an innovative sixteen-week curriculum for undergraduate translation majors focused on advanced digital resource construction and corpus processing. Theoretically grounded in a synthesis of the Intelligent-TPACK (AI-TPACK) framework and the Production-Oriented Approach (POA), the course challenges traditional instructional sequencing by immersing students in localized translation deficits and the chaotic realities of data harvesting, equipping them with inductive mastery of data cleaning, tokenization, and cross-lingual sentence alignment. The pedagogical design moves beyond mere technical vocationalism by positioning students as active algorithmic auditors. By integrating critical discourse analysis into the human-computer interactive workflow, students are trained to explicitly interrogate, document, and mediate the linguistic hegemony and systemic biases inherent in commercial AI models. Employing a rigorous, mixed-methods evaluation utilizing Multidimensional Quality Metrics (MQM), this paper demonstrates how shifting the learner's identity from a passive end-user to a critical language data architect cultivates the technological fluency, ethical grounding, and psychological resilience demanded by the contemporary language services industry.

Keywords: Translation Pedagogy, Corpus Processing, Generative Artificial Intelligence, Digital Resource Construction

1. Introduction

The landscape of translation studies, language service provision, and applied linguistics has undergone a profound epistemological and methodological transformation over the past decade. Driven by the exponential advancement of artificial intelligence (AI), neural machine translation (NMT), and large language models (LLMs), the traditional boundaries of the translation profession have irrevocably expanded (Cui *et al.*, 2025) [3]. The contemporary translation market demands a sophisticated, iterative synergy between human linguistic intuition and algorithmic processing power (Aleedy *et al.*, 2025) [1]. Consequently, translation education is compelled to transition from conventional skills-based bilingual training to multidimensional technological empowerment. This paradigm shift places a premium on digital resource construction, advanced corpus processing, and human-computer interactive workflows.

Within the context of Chinese higher education, this technological imperative intersects with the strategic macro-policy of the "New Liberal Arts" initiative. This national educational reform advocates for the deep, structural integration of traditional humanities disciplines with modern digital technologies, aiming to cultivate interdisciplinary, applied, and innovative talents capable of facilitating global cultural communication and participating in the digital economy (Pan & Cai, 2026) [6]. Against this background, this article presents a course design for "Artificial Intelligence and Corpus Processing," specifically for year-three undergraduate translation majors at a university in South China. The proposed curriculum systematically integrates the construction of monolingual and bilingual parallel corpora with advanced AI-assisted processing techniques. By positioning students as active auditors of digital language resources, this paper seeks to address the deficit in technological literacy that often characterizes undergraduate language/linguistics programs. The ensuing sections will delineate the literature underpinning this shift, outline the overarching theoretical frameworks governing the pedagogy, detail the design of the sixteen-week curriculum, and establish a rigorous mixed-methods empirical evaluation framework to measure instructional

efficacy.

2. Literature Review

To contextualize the proposed curriculum, it is necessary to synthesize the existing literature across three intersecting domains: the evolution of translation technology training, the methodologies of digital resource construction in translation studies, and the pedagogical paradigms governing technology-enhanced language learning.

2.1 The Evolution of Translation Technology Training

The integration of technology into translator training has historically followed the developmental trajectory of the language services industry. Early pedagogical interventions focused heavily on Computer-Assisted Translation (CAT) tools, primarily emphasizing translation memory (TM) management and basic terminology extraction. Empirical surveys conducted in the late 2010s indicated that while a majority of accredited translation programs offered technology courses, instruction was overwhelmingly dominated by traditional CAT software, often neglecting the underlying mechanics of machine translation or broader data literacy (Yan & Wang, 2022) [9]. This resulted in a facade of technological exuberance that masked dated instructional methods and an underestimation of the true value of computational linguistics.

The advent of deep learning and generative AI has disrupted this model. AI-enhanced systems now automate routine cognitive tasks, shifting the professional translator's role from generative drafting to complex post-editing, quality assurance, and cultural mediation. Current literature emphasizes the urgent need to bridge the gap between AI literacy, defined as a foundational understanding of AI concepts, neural networks, and ethical implications, and AI competency, which denotes the practical, reflective application of these tools in professional workflows. Scholars note that without this updated technological pedagogy, students risk developing a superficial understanding of language processing, leading to overreliance on automated tools, vulnerability to algorithmic bias, and a diminished capacity for critical, independent thought (Yang *et al.*, 2026) [10].

2.2 Digital Resource Construction and Corpus Processing

Corpus-based translation studies (CBTS) has matured into a robust empirical discipline over the past three decades, utilizing large collections of authentic texts to investigate translation universals such as explicitation, normalization, and simplification (Laviosa, 2021) [4]. However, traditional CBTS pedagogy often relegated students to the role of end-users, querying static, pre-compiled corpora using standard concordance software. The current technological frontier demands that students become capable of "digital resource construction" (Zhang *et al.*, 2023) [11].

In contemporary translation studies, particularly within the Chinese academic context, digital resource construction encompasses the entire process of data curation: from web crawling and optical character recognition (OCR) to data cleaning, tokenization, semantic annotation, and sentence-level alignment (Nguyen *et al.*, 2025) [5]. The literature increasingly highlights the utilization of AI to streamline this pipeline. For instance, projects like the Saudi Learner Translation Corpus (SauLTC) have demonstrated the efficacy of using Large Language Models (LLMs) in conjunction with multilingual sentence embedding models (such as Language-agnostic BERT Sentence Embedding, or LaBSE) to automate parallel sentence generation and compute cosine similarity for highly accurate cross-lingual alignment (Aleedy *et al.*, 2025) [1]. The ability to construct bespoke, domain-specific parallel corpora is recognized as a critical competency, enabling translators to train custom NMT engines and manage terminology in low-resource language pairs (Wang, 2026) [7].

2.3 Pedagogical Paradigms: DDL, AMRIL, and POA

The integration of corpus processing into the classroom is theoretically anchored in Data-Driven Learning (DDL). Pioneered in the early 1990s, DDL positions the learner as a "language explorer" who inductively discovers lexico-grammatical rules and pragmatic conventions by analyzing authentic corpus data (Wang, 2026) [7]. While traditionally a highly effective method for fostering learner autonomy, DDL in the era of generative AI has evolved into AI-Mediated Reflective Intercultural Learning (AMRIL). The AMRIL framework views AI not as an infallible oracle, but as a mediational agent within the student's Zone of Proximal Development. Through iterative cycles of data-driven noticing, AI-supported reflection, and communicative reconstruction, students use AI to externalize implicit cultural cues and semantic prosody, thereby developing deeper intercultural communicative competence.

Furthermore, the implementation of these technologies within the Chinese higher education system is highly compatible with the Production-Oriented Approach (POA) (Wen, 2018) [8]. POA directly challenges the traditional input-output sequencing of language education. Rather than front-loading theoretical abstractions, this curriculum initiates learning by confronting students with an immediate, high-stakes localized translation crisis such as an unedited, hallucination-riddled machine translation of a regional maritime law document or a culturally dense heritage text. By demanding an immediate diagnostic intervention, POA ensures that theoretical knowledge regarding syntax, alignment, and tokenization is acquired through the practical necessity of resolving algorithmic failures. This operationalizes the New Liberal Arts mandate to cultivate pragmatic, critical, and industry-ready graduates capable of navigating complex socio-technical deficits.

3. Theoretical Framework Governing the Curriculum

To ensure the pedagogical efficacy, academic rigor, and structural coherence of the proposed course, the curriculum design is anchored in a synthesis of advanced educational technology theories. The primary theoretical engine driving this course is the transition from traditional integration models to frameworks specifically adapted for generative, intelligent systems.

The Technological Pedagogical Content Knowledge (TPACK) framework has served as the dominant paradigm for conceptualizing how educators integrate digital tools into their teaching (Chiu, 2026) [2]. The original TPACK model posited that effective technology integration occurs at the intersection of Technological Knowledge (TK), Pedagogical Knowledge (PK), and Content Knowledge (CK). However, empirical research in the era of generative AI has exposed critical inadequacies in this foundational model. Traditional TPACK was developed during a period of non-generative,

relatively passive technology (e.g., interactive whiteboards, basic productivity software). Generative AI, by contrast, is an agentic, socio-ethically complex entity capable of autonomous content generation, deep data analysis, and simulated reasoning, which challenges the foundational assumptions of traditional technological integration.

To rectify this theoretical deficit, the proposed course design adopts the Intelligent-TPACK (AI-TPACK) framework (Chiu, 2026) [2]. This expanded model explicitly incorporates AI literacy and competency into the pedagogical matrix. Within the AI-TPACK model, Pedagogical Knowledge is expanded to AI-Pedagogical Knowledge (AI-PK), requiring an understanding of how generative systems adapt, make decisions, and influence the cognitive processes of learners. Technological Knowledge evolves into Intelligent Technological Knowledge (AI-TK), demanding fluency in prompt engineering, algorithmic auditing, and an understanding of neural network architectures.

Most crucially for a course on corpus processing, Content Knowledge intersects with AI to form Intelligent Technological Content Knowledge (AI-TCK). In the context of translation studies, AI-TCK dictates that students and instructors must understand not only the linguistic theories underpinning corpus design but also how AI algorithms interpret semantic relationships, execute word embeddings, and calculate cross-lingual alignments. This framework ensures that the use of LLMs for data cleaning or parallel sentence generation is approached critically, embedding ethical considerations (Ethics-K) directly into the workflow to combat algorithmic bias and data sovereignty issues.

4. Contextualizing the Learner Profile and Institutional Mandate

The successful implementation of an advanced AI-TPACK curriculum requires a precise calibration to the target demographic and the specific institutional environment. The course "Artificial Intelligence and Corpus Processing" targets year-three undergraduate translation majors at a university in South China.

By the inception of their third academic year, undergraduate translation majors have typically completed foundational coursework in bilingual proficiency, comparative linguistics, and basic translation theories. They possess a working knowledge of syntax, semantics, and cross-cultural communication protocols. However, traditional undergraduate curricula frequently exhibit a critical lacuna regarding computational linguistics, advanced terminology management, and large-scale data processing capabilities.

Year-three serves as the pivotal developmental juncture where students must transition from academic language learners into pre-professional language service providers and digital knowledge workers. At this stage, students possess the cognitive maturity necessary to engage with complex epistemological debates regarding AI's disruptive role in the translation industry. Yet, they also require highly structured, step-by-step technical scaffolding to master complex computational tools, such as utilizing regular expressions for text cleaning or running Python scripts for web crawling. The curriculum must therefore strike a delicate balance: providing rigorous technical instruction in corpus architecture while simultaneously fostering higher-order critical thinking regarding the ethical implications of data scraping, the homogenization effect of “translationese” in machine outputs, and the preservation of humanistic agency.

The university situated in this case study operates within the highly dynamic, rapidly evolving economic and technological ecosystem of the Guangdong-Hong Kong-Macao Greater Bay Area. This regional economy, characterized by intense cross-border e-commerce, international trade, and technological innovation, demands a workforce that is not only bilingual but also highly proficient in digital intelligence, data management, and AI-assisted workflows.

The proposed course design addresses this localized socioeconomic demand through its focus on digital resource construction. By training students to curate, clean, and align specialized parallel corpora relevant to regional industries such as international maritime law, regional cultural exports, or biomedical technology, the curriculum fulfills the institution's mandate to serve local economic development. Furthermore, this alignment with regional industry needs perfectly operationalizes the central tenets of the New Liberal Arts initiative, seamlessly blending humanities-based language study with empirical data science.

5. Architectural Blueprint of the Curriculum

The course is designed as an intensive, comprehensive sixteen-week module. The architecture of the curriculum reflects a carefully scaffolded pedagogical progression: initiating with theoretical foundations and AI literacy, transitioning into the mechanics of monolingual data processing, advancing to the complexities of bilingual parallel alignment, and culminating in advanced AI-assisted evaluation and professional post-editing workflows.

The following matrix delineates the thematic progression, the integration of specific technical tools, and the corresponding AI-TPACK competencies targeted within each instructional phase.

Module & Duration	Thematic Focus and Core Activities	Core Technical Tools & Platforms	AI-TPACK Competency Target
Module 1 (Weeks 1-4)	**Foundations of Corpus Linguistics & AI Literacy:** History of empirical approaches, corpus typologies, introduction to online platforms, and the epistemology of GenAI.	BNC, COCA, Sketch Engine, ChatGPT/Claude (for prompt engineering basics).	**AI-PK & TK:** Grasping the epistemological shift in empirical language research; understanding basic AI interaction, limitations, and prompt structures.
Module 2 (Weeks 5-7)	**Monolingual Digital Resource Construction:** Web crawling, Optical Character Recognition (OCR), data cleaning, tokenization, annotation, and frequency/collocation analysis.	AntConc, WordSmith, Python (basic scraping scripts), LLMs (for automated noise reduction and metadata generation).	**TCK & AI-CK:** Applying computational methods to extract linguistic patterns; evaluating AI efficacy in automating text normalization and domain classification.
Module 3 (Weeks 8-12)	**Bilingual Parallel Corpus Engineering:** Corpus translation studies (CTS) theory, sentence-level alignment techniques, managing cross-lingual structural divergence, and building parallel data architecture.	ParaConc, LF Aligner, ParaCLEAN pipeline, AI-driven sentence embedding models (e.g., LaBSE), LLMs for sentence generation.	**Intelligent-TPACK:** Mastering cross-lingual semantic mapping; empirically evaluating AI algorithms for structural and lexical correspondence; solving bilingual data scarcity.
Module 4 (Weeks 13-16)	**Corpus-Assisted Translation Studies (CATS) & MTPE:** Analyzing translated texts (political, technical), integrating corpora into CAT environments, human-AI collaborative post-editing.	SDL Trados/Phrase (MTPE environments), AI post-editing evaluation rubrics (MQM framework).	**Ethics-K & PCK:** Critically auditing machine translation output; resolving contextual discrepancies using bespoke corpora; internalizing professional ethical standards regarding AI use.

Module 1: Theoretical Foundations and AI Readiness (Weeks 1-4)

The initial phase of the curriculum is dedicated to establishing the empirical research paradigm that underpins modern corpus linguistics. Instruction begins by demystifying the concept of a corpus, tracing its historical evolution from early manual compilations to contemporary multi-billion-word digital repositories. Students are introduced to the critical theoretical distinctions between corpus-based (using corpora to validate existing theories), corpus-driven (deriving new theories directly from corpus data), and corpus-assisted methodologies.

This module introduces AI literacy as a core component of modern linguistic analysis. Students engage extensively with established, authoritative online platforms such as the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), and Sketch Engine. Through guided inquiry, they learn to execute complex searches to analyze word frequency, concordances, semantic prosody, and collocation. Meanwhile, they are trained in foundational prompt engineering. They learn how to structure queries to LLMs (such as ChatGPT or Claude) to simulate corpus analyses, extract linguistic patterns, and critically, to recognize the pervasive phenomena of algorithmic hallucinations, statistical probability masking as truth, and inherent data bias. This dual exposure ensures students understand the difference between retrieving empirically verified data from a structured corpus and generating probabilistic text via an LLM.

Module 2: Monolingual Digital Resource Construction (Weeks 5-7)

Transitioning from passive data analysis to active resource creation, the second module immerses students in the rigorous lifecycle of digital resource construction. Here, students learn the precise technical pipelines required to build a specialized, domain-specific monolingual corpus from unstructured raw data.

The instructional sequence begins with data harvesting techniques, introducing students to the ethical and technical parameters of web crawling, as well as the use of Optical Character Recognition (OCR) software for digitizing physical texts or archived PDF documents. Once raw data is acquired, the curriculum focuses heavily on data cleaning and normalization. Students learn to use regular expressions to remove HTML tags, boilerplate text, and non-textual artifacts.

The pedagogical design explicitly contrasts traditional manual or semi-automated data cleaning with AI-assisted preprocessing workflows. Students utilize established offline analytical tools such as AntConc and WordSmith to perform standard operations like concordancing, keyword extraction, and dispersion plotting on their newly cleaned data. Subsequently, they employ generative AI models to automate more complex tasks, such as generating metadata tags, classifying text domains, identifying specialized terminology, and performing initial semantic annotations. This comparative, hands-on approach allows students to empirically assess the efficiency gains, potential accuracy trade-offs, and systemic vulnerabilities of integrating AI into the corpus architecture pipeline.

Module 3: Bilingual Parallel Corpus Engineering (Weeks 8-12)

The most technically demanding and conceptually critical component of the curriculum lies in the construction of bilingual parallel corpora. Parallel corpora serve as the foundational bedrock for both descriptive translation studies and the training/fine-tuning of Neural Machine Translation (NMT) systems. The module begins with a deep dive into the history and rationale of Corpus Translation Studies (CTS), examining how the analysis of parallel data reveals translation universals—such as the tendency of translated texts to exhibit explicitation, normalization, and a reduction in lexical diversity compared to non-translated native texts.

The practical, laboratory-based component focuses on the intricate, computationally heavy process of bilingual text alignment. Students are trained on standard offline alignment and concordance tools, such as ParaConc and LF Aligner, to manage basic bilingual data analysis and resolve

simple 1-to-1 sentence alignments. However, the curriculum rapidly advances to introduce state-of-the-art AI-driven alignment methodologies. Drawing empirical inspiration from large-scale initiatives like the Saudi Learner Translation Corpus (SauLTC) and the ParaCLEAN pipeline, students learn how LLMs can be utilized to transform unstructured bilingual texts into highly accurate parallel sentence corpora.

Students explore the application of multilingual sentence embedding models (such as Language-agnostic BERT Sentence Embedding, or LaBSE). They learn how these models convert sentences into high-dimensional vectors and compute cosine similarity to automate the identification of semantic equivalence across languages, even in the presence of severe cross-lingual structural divergence. This module demystifies the "black box" of machine translation, equipping translation majors with the sophisticated NLP skills required to curate, clean, and align high-quality training datasets, an invaluable skill in the context of low-resource language domains or highly specialized industrial translation tasks.

Module 4: Applied CATS and Human-AI Collaborative Workflows (Weeks 13-16)

The culminating module of the curriculum synthesizes corpus construction with practical, industry-standard translation execution and critical academic analysis. Students engage in rigorous paper reading sessions, focusing on peer-reviewed corpus-assisted analyses of translated texts across diverse domains, including political discourse, specialized applied translation (e.g., medical or legal), and interpreting transcripts.

A significant pedagogical focus is placed on advanced Machine Translation Post-Editing (MTPE). The curriculum adopts a "guidance-based" approach to MTPE, wherein students are explicitly trained not to passively accept machine-generated outputs. Instead, they engage in an iterative, dialogic interaction with generative AI tools. Students integrate the bespoke bilingual corpora they constructed in Module 3 directly into professional CAT environments (such as SDL Trados or Phrase). They utilize their own digital resources to verify highly specialized terminology, check semantic prosody, and ensure stylistic consistency, using empirical corpus data to audit, refine, and elevate AI-generated draft translations. This instructional design cultivates the specific cognitive profile required for modern language professionals, shifting the student's primary effort away from raw lexical generation and toward high-level cultural mediation, stylistic refinement, and authoritative quality assurance.

6. Pedagogical Mechanisms and Human-Computer Interaction

The integration of digital resource construction into the translation curriculum is conceptualized not merely as a technical, vocational exercise, but as a profound pedagogical mechanism for enhancing deeper translation competence, metacognitive awareness, and digital literacy. The arduous process of building a corpus enables students to confront the micro-linguistic realities of morphology and syntax, as well as the macro-textual realities of discourse and pragmatics, in ways that abstract theoretical study cannot replicate.

6.1 Institutionalizing the "Human-in-the-Loop" Workflow

The curriculum advocates strongly for a "human-in-the-loop" approach to corpus processing and translation generation. While modern AI algorithms possess an unprecedented, superhuman capacity for processing vast datasets, rapidly identifying statistical collocations, and suggesting probable alignments, they inherently lack the socio-cultural context, historical awareness, and ethical discernment required for nuanced linguistic analysis. Therefore, the pedagogical design mandates that students act as expert validators and algorithmic auditors throughout the curriculum.

For example, when a student utilizes an LLM to generate a bilingual terminology glossary from a crawled maritime law dataset, the AI will inevitably produce false positives or contextually inappropriate alignments due to polysemy, syntactic ambiguity, or cultural divergence. The student must apply their specialized domain knowledge and human linguistic intuition to manually review, filter, and refine the AI's output. This constant, iterative cycle of machine generation followed by critical human curation not only enhances the empirical quality of the resulting digital resource but dramatically improves the student's metacognitive awareness of translation strategies and structural differences between English and Chinese.

6.2 Managing Cognitive Load and Interaction Complexity

While AI integration significantly increases overall translation efficiency and data processing speed, the curriculum design explicitly acknowledges the cognitive paradox of AI in translation workflows: while offering immense conceptual support, AI frequently introduces entirely new technical and cognitive burdens.

Empirical research utilizing advanced tracking methodologies reveals that AI-assisted post-editing (AIPE) can severely increase interaction complexity. Students in an AIPE environment must simultaneously divide their visual and cognitive attention across multiple interfaces: the source text, the initial machine translation output, the traditional CAT tool interface, specialized corpus concordance windows, and dynamic LLM prompt interfaces. This constant cognitive context-switching can lead to cognitive overload, resulting in erratic pause durations, increased anxiety, and variable editing efficiency.

To mitigate this cognitive strain, the course design incorporates deliberate, structured scaffolding. Initial tasks in Module 2 involve highly constrained, pre-cleaned corpora and specific, rule-based alignment exercises. As students develop technical fluency and automaticity with the interfaces, the parameters are gradually relaxed in Modules 3 and 4, introducing larger, noisier datasets and open-ended generative AI interactions. This progressive complexity ensures that students develop the psychological resilience and technical agility necessary to manage sophisticated, multi-agent LLM workflows without becoming overwhelmed by the interface demands.

7. Empirical Evaluation Framework for Course Effectiveness

To ensure the theoretical soundness, pedagogical validity, and practical impact of this highly innovative course design, it is imperative to implement a rigorous, empirical evaluation framework. Moving beyond superficial, self-reported metrics of student satisfaction, the evaluation must capture deep cognitive changes, behavioral adaptations in workflow, and the actual, objective quality of the digital artifacts produced by the students. The proposed evaluation strategy utilizes a convergent mixed-methods design grounded in the New World Kirkpatrick Model.

7.1 The Modified Kirkpatrick Evaluation Matrix

The evaluation framework assesses the pedagogical intervention across four distinct, hierarchical levels, employing both quantitative statistical analysis and qualitative thematic inquiry to triangulate the findings. The structured approach guarantees that both the process of learning and the final products are rigorously assessed.

Evaluation Level	Target Construct	Assessment Methodology and Empirical Instruments
Level 1: Reaction	Learner engagement, perceived relevance, psychological safety, and initial cognitive load.	Post-module digital surveys utilizing 5-point Likert scales to measure the perceived utility of specific tools (e.g., ParaConc vs. LLM sentence alignment). Qualitative focus groups to assess the psychological safety of the AI-mediated learning environment.
Level 2: Learning	Acquisition of corpus linguistics concepts, prompt engineering skills, and AI-TPACK development.	Pre- and post-course knowledge tests assessing empirical linguistics theory. Administration of a validated AI-TPACK scale to measure quantitative shifts in Technological Knowledge (TK) and AI-Pedagogical Knowledge (AI-PK). Evaluation of midterm digital resource metadata schemas.
Level 3: Behavior	Application of corpus tools in authentic translation workflows; adaptation of human-AI collaboration strategies.	Analysis of key-logging, screen-recording, and eye-tracking data (where lab facilities permit) during MTPE tasks to quantify interaction complexity, fixation durations, saccade amplitudes, and reliance on AI prompts versus traditional corpus queries.
Level 4: Results	Quality of the final digital artifacts (parallel corpora) and the accuracy of context-aware translation outputs.	Expert instructor evaluation of the student-built parallel corpora using the Multidimensional Quality Metrics (MQM) framework. Assessment criteria include alignment accuracy, noise reduction efficiency, and terminology consistency across the corpus.

7.2 Advanced Metrics: Context-Aware Evaluation and MQM Integration

A pervasive issue in the evaluation of traditional translation technology courses is the reliance on isolated, sentence-level metrics (such as BLEU or METEOR scores). These metrics can be highly misleading, often producing high scores for segments that appear grammatically correct in isolation but fail to maintain logical cohesion, register, or consistent terminology within the broader discourse of a complete document. To counteract this limitation, the evaluation of this course prioritizes document-level, context-aware assessment methodologies.

The final summative assessment—a comprehensive group project accounting for 20% of the total course grade—requires students to submit a fully functional, aligned, and annotated bilingual corpus, accompanied by a reflective analytical report. The evaluation of this project employs a customized Multidimensional Quality Metrics (MQM) rubric. This rubric is specifically tailored to identify errors in corpus architecture, such as sentence mispairing, data omission, under-confidence in machine output (resulting in unnecessary human edits), and failure to detect and correct algorithmic hallucinations.

Furthermore, reflecting the cutting-edge nature of the curriculum, the evaluation process incorporates AI as a secondary, automated assessor. Utilizing advanced LLMs (such as ChatGPT-4o) prompted with the exact same MQM rubrics used by the human instructors, the course allows for a fascinating comparative analysis between human grading and AI-generated diagnostic feedback. Recent empirical studies indicate a remarkably high level of statistical agreement (with no statistically significant differences in paired sample t-tests) between expert translation instructors and advanced LLMs in identifying translation errors and evaluating target-language quality. By exposing students to this dual-evaluation mechanism, the course not only provides immediate, highly detailed feedback but also provokes vital critical discussions regarding the reliability, reasoning processes, transparency, and potential biases inherent in automated quality assessment systems.

8. Broader Socio-Ethical Implications for Translation Studies

The integration of artificial intelligence and advanced corpus processing into the undergraduate translation curriculum extends far beyond the mere acquisition of technical, vocational skills. It catalyzes a fundamental re-evaluation of translator agency, professional identity, and the ethical responsibilities of language workers in the digital age.

As students actively harvest data from the open internet to construct their corpora, they are invariably exposed to the systemic biases, cultural stereotypes, and linguistic hegemony embedded within global digital discourse. LLMs and NMT systems, which are trained on vast, largely unfiltered datasets, frequently reproduce these hegemonic linguistic patterns, default to Western-centric cultural assumptions, and exhibit documented gender and racial biases. A critical, non-negotiable component of the AI-TPACK framework, specifically located within the Ethics-K (Ethical Knowledge) domain—involves training students to explicitly recognize, document, and mitigate these biases during the data cleaning, annotation, and alignment phases.

The curriculum mandates that students critically evaluate the demographic and cultural representativeness of their self-compiled corpora. By engaging in the deliberate curation of specialized, historically marginalized texts, such as regional Lingnan cultural narratives, Hakka language data, or specific minority discourses, students actively participate in counteracting the homogenization of language driven by dominant commercial AI models. This rigorous process fosters a sophisticated level of civic and media literacy,

effectively transforming students from passive operators of translation software into ethical stewards of cross-cultural digital resources.

The rapid proliferation of generative AI has understandably provoked significant anxiety among translation trainees regarding impending job displacement, market devaluation, and the obsolescence of human linguistic expertise. This course design addresses these valid affective concerns directly by fundamentally reframing the professional identity of the future translator.

Through the rigorous, hands-on practice of corpus construction, data alignment, and LLM orchestration, students arrive at an empirical realization: while AI excels at rapid lexical generation and probabilistic pattern matching, it fundamentally lacks the capacity for genuine communicative intent, empathetic cultural mediation, and strategic, context-aware decision-making. The curriculum demonstrates practically that the most highly valued skills in the contemporary language services market are no longer the rote memorization of bilingual vocabulary, but rather advanced data management, algorithmic auditing, and high-level post-editing.

By mastering the underlying architecture of parallel corpora, students position themselves not in futile competition with AI, but as the indispensable domain experts who manage, train, correct, and refine the AI systems. This strategic shift in professional identity, from traditional "bilingual translator" to "language data architect and cultural mediator", is crucial for ensuring the long-term employability, relevance, and psychological resilience of graduates as they enter the highly automated language services industry.

9. Conclusion

The intersection of artificial intelligence, corpus linguistics, and translation pedagogy represents one of the most critical and complex frontiers in contemporary higher education. The proposed course design for "Artificial Intelligence and Corpus Processing" provides a blueprint for navigating this transitional era. By explicitly grounding the curriculum in the advanced AI-TPACK framework and the Production-Oriented Approach, the course transcends the limitations of traditional, passive software training. Instead, it immerses year-three translation majors in the complex, authentic, and highly technical processes of monolingual and bilingual digital resource construction.

Through a meticulously scaffolded progression, moving from basic corpus querying and theoretical epistemology to advanced LLM-assisted parallel sentence alignment and context-aware post-editing, the curriculum equips students with the advanced computational proficiencies demanded by the modern language services market. Crucially, it couples this intensive technical training with profound ethical inquiry and critical reflection, ensuring that students possess the analytical tools necessary to navigate algorithmic biases and maintain humanistic agency in AI-mediated workflows. Supported by a robust, mixed-methods empirical evaluation framework utilizing the Kirkpatrick model and MQM rubrics, this course design not only fulfills the interdisciplinary mandates of China's New Liberal Arts initiative but also establishes a forward-looking, highly replicable paradigm for cultivating the next generation of linguistically adept, technologically empowered, and ethically grounded translation professionals.

10. Acknowledgments

This work was supported by the Higher Education Teaching Reform Project of Guangdong University of Education (Grant No. JXGG241248) and the Guangdong Educational Science Planning Project (Higher Education Special) (Grant No. 2024GXJK636).

11. References

1. Aleedy M, Alshihri F, Meshoul S, Al-Harthi M, Alramlawi S, Aldaihani B, *et al*. Designing AI-powered translation education tools: A framework for parallel sentence generation using SauLTC and LLMs. PeerJ Computer Science. 2025; 11:e2788. Doi: https://doi.org/10.7717/peerj-cs.2788
2. Chiu TK. Intelligent-TPACK (I-TPACK) framework developed from TPACK through integration of artificial intelligence literacy and competency. Interactive Learning Environments, 2026, 1-16.
3. Cui F, Li D, Zhuang C. Introduction: Transforming translation education through Artificial Intelligence. The Interpreter and Translator Trainer. 2025; 19(3-4):227-233.
4. Laviosa S. Corpus-based Translation Studies: Theory, Findings, Applications (Vol. 17). The Netherlands, Brill, 2021.
5. Nguyen TNN, Tran TT, Nguyen NHA, Lam HP, Nguyen HMS, Tran NAT. The Benefits and Challenges of AI Translation Tools in Translation Education at the Tertiary Level: A Systematic Review. International Journal of TESOL & Education. 2025; 5(2):132-148. Doi: https://doi.org/10.54855/ijte.25527
6. Pan Y, Cai P. Research on the Integration Path of Innovation and Entrepreneurship Education and Foreign Language Major Education under the Background of "New Liberal Arts". Open Access Library Journal. 2026; 13:e14947. Doi: https://doi.org/10.4236/oalib.1114947
7. Wang M. AI-Enhanced Corpus-Driven Pedagogy for Intercultural Communicative Competence Development: A Theoretical Model and Feasibility Study. English Language Teaching. 2026; 19(1). Doi: 10.5539/elt.v19n1p50
8. Wen Q. The production-oriented approach to teaching university students English in China. Language Teaching. 2018; 51(4):526-540.
9. Yan D, Wang J. Teaching data science to undergraduate translation trainees: Pilot evaluation of a task-based course. Frontiers in Psychology. 2022; 13:939689.
10. Yang C, Hou S, Zhao M, Yan J, Chen J. Translation students' perceptions of the integration of artificial intelligence in translation education: A constructivist approach. Artificial Intelligence in Education. 2026; 2(2):157-174.
11. Zhang H, Lu R, Liu P. Constructing a Model of Intelligent Learning Space for Vocational Education in the Age of Intelligence. In: Proceedings of the 2023 4th International Conference on Artificial Intelligence and Education (ICAIE 2023). the Netherlands, Atlantis Press, 2023, 137-152.