ISSN: 2386-9658 https://doi.org/10.21071/arf.v37i.19074

BThe Convergence of Corpus Linguistics and Generative AI: Reshaping Translation, Education, and Specialized Communication

(Convergencia entre Lingüística de Corpus e IA Generativa: una restructuración de traducción, B educación y comunicación especializada)

LETICIA MORENO-PÉREZ
https://orcid.org/0000-0001-9211-7166
leticia.moreno@uva.es
Universidad de Valladolid https://ror.org/01fvbaw18

BELÉN LÓPEZ-ARROYO
https://orcid.org/0000-0002-9171-1910
mariabelen.lopez@uva.es
Universidad de Valladolid https://ror.org/01fvbaw18

Fecha de recepción: 1 de febrero de 2026
Fecha de aceptación: 23 de febrero de 2026

Abstract: The rapid integration of Large Language Models (LLMs) has catalyzed a paradigm shift in applied linguistics. We explore the synergy between traditional corpus methodologies and artificial intelligence across three core dimensions: resources, practice, and pedagogy. Through a narrative review of nine recent studies presented in this volume, this paper proposes a model for GenAI–corpus integration, moving from manual linguistic labor toward augmented linguistic workflows. We examine how AI automates the compilation of specialized medical corpora, the linguistic monitoring of synthetic text, and the identification of systemic gender biases and cultural mistranslations in machine output. Finally, the literacies required for the next generation of linguists are outlined. This work is intended for researchers, practitioners, and teacher educators seeking a structured approach to AI implementation in translation and language education.

Keywords: Generative AI. Corpus Linguistics. Translation. Education. Specialized Communication.

Resumen: La rápida integración de los Grandes Modelos de Lenguaje (LLMs por sus siglas en inglés) ha generado un cambio de paradigma en la lingüística aplicada. En el presente artículo exploramos la sinergia entre las metodologías de corpus tradicionales y la inteligencia artificial en tres dimensiones esenciales: recursos, práctica y pedagogía. A través de una revisión narrativa de nueve estudios recientes presentados en este volumen, este artículo propone un modelo de integración entre la IA generativa y los corpus, avanzando desde el trabajo lingüístico manual hacia flujos de trabajo lingüísticos asistidos por IA. Examinamos de qué manera la IA automatiza la compilación de corpus médicos especializados, la monitorización lingüística de texto sintético y la identificación de sesgos de género sistémicos y errores de traducción cultural en resultados generados por máquinas. Finalmente, se describen las competencias necesarias para la próxima generación de lingüistas. Este trabajo está dirigido a investigadores, profesionales y formadores de docentes que buscan un enfoque estructurado para la implementación de la IA en la traducción y la educación en el ámbito de la lingüística.

Palabras clave: IA generativa. Lingüística de corpus. Traducción. Educación. Comunicación especializada.

1. Introduction: The Post-2022 Landscape

The landscape of language services and applied linguistics has undergone a seismic shift since the late 2022 widespread adoption of Large Language Models. To understand this transition, one must contrast it with the pre-LLM era. Previously, specialized communication relied on Rule-Based or Statistical Machine Translation (SMT) and manual corpus exploitation via concordancers. Today, as Lefer and Bodart (2026) observe, translation "from scratch" has been superseded by a hybrid model of Machine Translation Post-Editing (MTPE), now prevalent in approximately half of all professional projects.

This technological surge has not only altered how we translate but has also redefined the creation of linguistic resources and the methodologies of language education. We are moving from "assisted translation" to "augmented linguistics", where the boundary between human and machine authorship is increasingly porous. In this sense, three central questions need to be addressed:

1. How are corpus compilation and monitoring practices being reconfigured by AI automation?
2. Which linguistic and ethical risks—specifically regarding culturemes and gender—emerge in human–AI collaborative workflows?

3. What specific technical and critical competences must future professionals and educators acquire to navigate this transition?

2. Evolution of Corpus Resources: Automation and Monitoring

The synergy between traditional corpus linguistics and AI is most evident in the development of specialized resources, shifting the linguist's role from data entry to algorithmic oversight.

2.1. Scaling Specialized Corpora

In the medical domain, Sánchez Nieto (2026) details the creation of the Merck DE-ES corpus, a 12-million-word parallel resource. By utilizing scripts developed with AI models like Copilot and LeChat, the project automated the extraction and alignment of 400,000 bisegments. This automation drastically reduces project timelines, making the creation of large-scale, register-specific corpora (expert-to-expert vs. expert-to-layperson) feasible for smaller research teams. Beyond compilation, this resource enables high-precision terminology management, readability studies for patient communication, and comparative register analysis that was previously restricted by manual processing bottlenecks.

2.2. Monitoring the "Robotic" Voice

As synthetic text floods the digital ecosystem, Alonso Simón *et al.* (2026) introduce the ROBOT-TALK monitor corpus. This resource identifies "robotic markers"—specific lexical repetitions, lack of syntactic variety, and flattened discourse-pragmatic nuances—that differentiate LLM output from human authorship. These monitor corpora are vital for model evaluation and fine-tuning protocols; they provide a baseline for "naturalness" and help developers identify where models fail to replicate human discursive creativity in scientific and news genres.

3. Human–Machine Interaction: Domain-Specific and Systemic Risks

As AI becomes integrated into professional workflows, we must distinguish between risks inherent to specific domains and those that are systemic across all AI-mediated communication.

3.1. Domain-Specific Risks: The Case of Culturemes

In specialized gastronomy, Morales-Jiménez (2026) evaluates how systems handle "culturemes" (e.g., PDO vinegars). The research highlights a specific domain risk: the tendency of AI toward literalism or "convergent" techniques that preserve the source culture at the expense of target-audience

comprehension. Without expert human intervention, these models fail to adapt technical sheets for lay audiences, often resulting in cultural mistranslations that can impact the commercial and heritage value of the products.

3.2. Systemic Risks: Gender Bias and MTPEAS

Beyond domain errors lie systemic risks, such as gender bias propagation. García Serrano and Toledo-Báez (2026) demonstrate that NMT systems like DeepL often default to masculine archetypes, which are frequently "overlooked" by post-editors without specific training. To address this, Lefer and Bodart (2026) propose the Machine Translation Post-Editing Annotation System (MTPEAS). This framework uses concrete annotation categories—Adequacy, Fluency, Register, and Bias—to track student edits. By standardizing how bias and errors are tagged, MTPEAS allows educators to pinpoint exactly where human supervision fails, turning the post-editing process into a measurable data point for professional certification.

The opacity of AI integration in journalism is described by Scilabra and Russo (2026) analyzing gender bias not through overtly discriminatory terminology, but through what appears to be "neutral" scientific framing. With sentiment analysis tools their findings demonstrate that technical-scientific registers—precisely the linguistic mode that conveys objectivity and authority—can naturalize binary gender categories while obscuring the institutional actors responsible for discriminatory coverage. What distinguishes this study from earlier bias detection work is its insistence that the problem is not simply poor training data but the structural incompatibility between probabilistic language models and the context-dependent nuances required for equitable representation.

4. Pedagogical Integration: From Content to Literacy

The educational sphere is moving from using AI as a "search engine" to using it as a multifaceted teaching partner.

4.1. The Three Teaching Functions

Qi (2026) evaluates AI performance across three distinct functions using Prepositional Regime Complements (CRP) in Spanish:

- Content Explanation: Models excel at defining rules but often lack discourse-level context.
- Exercise Generation: AI can produce vast quantities of drill material but struggles with pedagogical sequencing.
- Feedback/Resolution: While accurate in solving, models often fail to explain *why* an error occurred in a way that aligns with student level.

4.2. A Framework for AI Literacy

Successful integration requires a tripartite literacy framework, as explored in the teacher-training studies by Badilla Ramos (2026) and the genre-based research of Labrador de la Cruz (2026):

- Technical Literacy: Mastering tools like *Grammarly* or *NoteGPT* for basic efficiency.
- Design Literacy: Using tools like *Notebook LM* to query specific, "grounded" corpora (e.g., cheese descriptions) to generate genre-accurate content without hallucinations.
- Critical Literacy: Identifying the ethical and social biases (as seen in the MTPE studies) inherent in these models.

5. Discussion: The Augmented Human-in-the-Loop Model

The integration of GenAI into linguistic workflows does not signal the displacement of the human professional but rather a reconfiguration of their cognitive load. We propose an Augmented Linguistic Workflow (ALW) model that categorizes tasks based on the level of "interpretive agency" required. While AI excels at pattern replication and high-volume processing, the human remains the primary agent for contextual mediation and ethical accountability.

5.1. Task Allocation and Agency

In this model, tasks are no longer binary (human vs. machine) but are distributed along a spectrum of autonomy.

- AI-Led Autonomy (Low Interpretive Agency): These are "brute force" tasks where efficiency is the primary goal. As demonstrated by Sánchez Nieto (2026), the use of AI scripts for the alignment of 12 million words in the *Merck DE-ES* corpus represents a task where the machine's ability to process massive datasets far exceeds human capability, with human intervention limited to spot-checking and parameter setting.
- Shared/Collaborative Agency (Medium Interpretive Agency): This involves "grounded" creativity. In Labrador's (2026) use of *Notebook LM*, the AI assists in extracting stylistic patterns from a specific cheese-description corpus, but the researcher must design the prompts and interpret the findings within the framework of genre analysis.
- Human-Led Sovereignty (High Interpretive Agency): These tasks involve social, ethical, and cultural stakes. As seen in the work of Morales-Jiménez (2026) and García Serrano & Toledo-Báez (2026), the detection of cultureme nuances and the mitigation of "invisible" gender biases require a level of world knowledge

and ethical sensitivity that LLMs—which function on probabilistic next-token prediction—cannot replicate.

5.2. Formalizing the ALW Model

The following table formalizes how the nine studies in this volume contribute to an integrated model of human-AI collaboration:

Workflow Dimension	AI Primary Contribution (Efficiency)	Human Primary Contribution (Quality & Ethics)	Key Reference
Corpus Engineering	Rapid data scraping, alignment of millions of words, metadata automation.	Defining registers (expert vs. lay), ensuring representativeness, cleaning noise.	Sánchez Nieto (2026); Alonso Simón et al. (2026); Scilabra & Ruso (2026)
Translation & Localization	Initial drafting, consistency across large volumes, register shifting.	Cultureme mediation, gender-neutral editing, linguistic "naturalness."	Morales-Jiménez (2026); Toledo-Báez & García Serrano (2026)
Pedagogy & Training	Content generation, exercise drafting, basic grammar explanation.	Pedagogical sequencing, empathy-based feedback, critical AI literacy.	Qi (2026); Badilla Ramos (2026); Labrador (2026)
Analytical Oversight	Processing edits at scale.	Categorizing post-editing effort and bias using frameworks like MTPEAS.	Lefer & Bodart (2026)

Table 1: Integrated model of human-AI collaboration. *Alfinge* (2026)

Conclusion and Recommendations

The integration of AI into linguistics is irreversible, but its success depends on strategic augmentation rather than total automation. To ensure an ethical and effective future, this paper offers three recommendations:

1. For Academic Journals: Establish reporting standards requiring authors to disclose the extent of AI use in corpus compilation and text generation.
2. For Curricula: Integrate mandatory "AI-and-Corpus" modules that focus on critical literacy and prompt engineering for specialized domains.
3. For Industry: Develop auditing routines for MTPE workflows to detect systemic biases before they reach the public sphere.

The most powerful tool in the digital age remains the critically trained human mind, now augmented by the massive processing power of Generative AI.

References

ALONSO SIMÓN, Lara; FERNÁNDEZ-PAMPILLÓN CESTEROS, Ana María, "El corpus ROBOT-TALK para el reconocimiento del origen robótico de textos en español". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 9-32.

BADILLA RAMOS, Karina, "Perceptions of Five Costa Rican EFL Teachers Regarding the Integration of Large Language Models: An Exploratory Study". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 33-49.

LABRADOR DE LA CRUZ, Belén, "Capitalizing on genre-based corpora with the use of the AI-powered research tool Notebook LM". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 51-75.

LEFER, Marie Aude; BODART, Romane, "Student post-editing corpora: Collection, annotation and analysis". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 77-96.

MORALES-JIMÉNEZ, Juan Pedro, "Análisis de las técnicas de traducción y adaptación de culturemas en el vinagre DOP español mediante IA y TAN". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 97-119.

QI, Wanyun, "Rendimiento de modelos de IA generativa en tres funciones docentes: análisis de tareas con CRP en ELE". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 121-141.

SÁNCHEZ NIETO, María Teresa, "*MSD-Manuals* DE-ES. Un corpus de comunicación especializada mediada médica en el Parallel Corpus of German and Spanish (PaGeS)". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 143-169.

SCILABRA, J. Mary; RUSSO, Dario, "Intelligenza Artificiale e Fake News: Analisi Linguistica, Gender Bias ed Etica nella comunicazione digitale dei media". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 171-193.

TOLEDO-BÁEZ, Cristina; GARCÍA SERRANO, María Jesús, "Sesgos invisibles: género y posedición en traducción automática neuronal". In: *Alfinge. Revista de Filología*, 37, 2026, pp. 195-215.