
# PaCor 2025

**5th International Symposium Parallel Corpora PaCor 2025**

**"(Parallel) Corpus-Based Approaches to Language and Data: Generative AI Applications in Focus"**

28-30 May 2025

University of Valladolid (Spain)


Universidad de Valladolid


Ayuntamiento de
Valladolid


UVa
IA
Centro de Inteligencia Artificial


Facultad de
Filosofía y Letras


FACULTAD DE COMERCIO
Universidad de Valladolid


Virt
UVa


CEA
Centro de Estudios de Asia
Universidad de Valladolid


nlpgo


Universidad de Valladolid
Departamento de Filología
Francesa y Alemana


Departamento de Filología Inglesa
UVa
Facultad de Filosofía y Letras


Universidad de Valladolid
Departamento de Lengua Española


PaCor 2025

# *COMITÉS / COMMITTEE*

## *Organizador* / Organizing committee

**Presidentas / Chairs**

Belén López Arroyo
Leticia Moreno Pérez

**Miembros / Members**

Susana Álvarez Álvarez
Esther Álvarez de la Fuente
Pilar Garcés García
Giovanna Mapelli
Francisco Javier Muñoz Acebes
Isabel Pizarro Sánchez
María Luisa Rodríguez Muñoz
Secretarías técnicas:
Tamara Gómez Carrero
Lucía Sanz Valdivieso
Qianting Yuan

**Colaboradores / Collaborators**

Alejandro Calleja
Irene Díez
Marcela Ferrándiz
Andrea Herguedas
Yolanda Infante
Borislava Ivanova
Aroa López
Jose Magallanes
Lucía Mañeru
Raquel Martínez
Iria Méndez
Susana Molinero
María Morchón

## Científico / Scientific committee

Susana Álvarez Álvarez (Universidad de Valladolid, España)
Esther Álvarez de la Fuente (Universidad de Valladolid, España)
Laura Filardo Llamas (Universidad de Valladolid, España)
Pilar Garcés García (Universidad de Valladolid, España)
Inés González (Universidad de León, España)
Suguru Ishizaki (Carnegie Mellon University, USA)
Marlen Izquierdo Fernández (Universidad del País Vasco, España)
David Kaufer (Carnegie Mellon University, USA)
Belén Labrador de la Cruz (Universidad de León, España)
Belén López Arroyo (Universidad de Valladolid, España)
Elizabeth Marshmann (University of Ottawa, Canada)
Rosario Martín Ruano (Universidad de Salamanca, España)
Leticia Moreno Pérez (Universidad de Valladolid, España)
Francisco Javier Muñoz Acebes (Universidad de Valladolid, España)
Giuseppe Palumbo (Universitá degli Studi di Trieste, Italia)
María Pérez Blanco (Universidad de León, España)
Carla Quinci (Universitá degli Studi di Padova, Italia)
Rosa Rabadán Álvarez (Universidad de León, España)
Noelia Ramón García (Universidad de León, España)
Maria Teresa Sánchez Nieto (Universidad de Valladolid, España)
Azahara Veroz (Universidad de Córdoba, España)
Xiaobo Wang (Sam Houston State University, USA)

# LIBRO DE RESÚMENES - BOOK OF ABSTRACTS

## TEMÁTICAS - TRACKS

1. Aplicaciones de la IA generativa en la traducción y los corpus multilingües / Generative AI applications in translation and multilingual corpora

2. Codificación de corpus / Corpus encoding

3. Consideraciones interlingüísticas e interculturales en textos generados por IA / Cross-linguistic and cross-cultural considerations in AI-generated texts

4. El papel de los modelos de lenguaje de gran escala (LLMs) en la lingüística de corpus / The role of large language models (LLMs) in corpus linguistics

5. Mejora de la enseñanza de idiomas con IA y corpus / Enhancing language teaching with AI and corpora

6. Traducción multilingüe y redacción profesional / Multilingual translation and professional writing

## CONFERENCIAS PLENARIAS / PLENARY LECTURES



## TALLER / WORKSHOP



## 1. APLICACIONES DE LA IA GENERATIVA EN LA TRADUCCIÓN Y LOS CORPUS MULTILINGÜES / GENERATIVE AI APPLICATIONS IN TRANSLATION AND MULTILINGUAL CORPORA



## 2. CODIFICACIÓN DE CORPUS / CORPUS ENCODING

## 3. CONSIDERACIONES INTERLINGÜÍSTICAS E INTERCULTURALES EN TEXTOS GENERADOS POR IA / CROSS-LINGUISTIC AND CROSS-CULTURAL CONSIDERATIONS IN AI-GENERATED TEXTS



## 4. EL PAPEL DE LOS MODELOS DE LENGUAJE DE GRAN ESCALA (LLMS) EN LA LINGÜÍSTICA DE CORPUS / THE ROLE OF LARGE LANGUAGE MODELS (LLMS) IN CORPUS LINGUISTICS

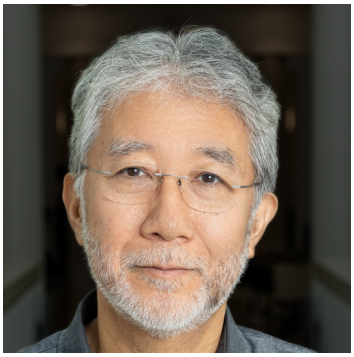


## 5. MEJORA DE LA ENSEÑANZA DE IDIOMAS CON IA Y CORPUS / ENHANCING LANGUAGE TEACHING WITH AI AND CORPORA



## 6. TRADUCCIÓN MULTILINGÜE Y REDACCIÓN PROFESIONAL / MULTILINGUAL TRANSLATION AND PROFESSIONAL WRITING

# Conferencias Plenarias / Plenary Lectures



**SUGURU ISHIZAKI** es catedrático de Inglés en la Carnegie Mellon University, Pittsburgh, Pensilvania, USA donde también es el director de los programas de escritura profesional y técnica. Su investigación se centra en el diseño de entornos de producción escrita con IA mejorada. Recientemente, ha coeditado, junto con Belén López-Arroyo y Belle Wang, el volumen especial de *IEEE Transactions on Professional Communication* journal, titulado *Building Bridges Between Technical and Professional Communication and Translation Studies*. Anteriormente a su posición actual, trabajó como ingeniero senior en Qualcomm donde se centraba en investigación, desarrollo y gestión de productos para aplicaciones móviles. También fue profesor en la Carnegie Mellon's School of Design en Pittsburgh, Pensilvania, USA. Obtuvo su título de máster y doctorando en MIT's Media Laboratorty y ha sido presidente de la IEEE Professional Society.

**SUGURU ISHIZAKI**, Ph.D., is a Full Professor of English at Carnegie Mellon University, where he directs its Professional & Technical Writing Programs. His research focuses on designing AI-enhanced writing environments and computer-assisted rhetorical analysis. He recently co-edited *Building Bridges Between Technical and Professional Communication and Translation Studies*, a special issue of *IEEE Transactions on Professional Communication*, with Belén López Arroyo and Belle Wang. Before his current role, he was a senior staff engineer at Qualcomm, working on research, development, and product management for early mobile applications. He previously served as a faculty member at Carnegie Mellon's School of Design. He earned his Master's degree and Ph.D. from MIT's Media Laboratory and is a past president of the IEEE Professional Communication Society.

## Future of writing in the discipline and professions

Writing is inherently complex, demanding significant cognitive effort across multiple dimensions: generating and organizing ideas, structuring coherent arguments, and refining expression to effectively communicate with readers (Flower & Hayes, 1980, 1981). While technological advancements have progressively eased the mechanical aspects of writing over the years, from word processors to grammar checkers, the emergence of generative AI presents unprecedented possibilities to support writers throughout their creative process. However, this transformative potential comes with notable risks, particularly the danger of overreliance on AI systems, which could potentially undermine the development of critical thinking skills and creativity —fundamental elements that distinguish effective writing.

This plenary talk will present a comprehensive vision for a human-AI partnership that enhances the writing process while preserving its essential human elements. By strategically reducing cognitive load, extending prewriting activities through ideation support, leveraging sophisticated genre knowledge, and improving the review and revision process, AI can empower writers to focus more intensively on higher-order thinking and creative expression (Kaufer & Ishizaki 2024). This vision will be illustrated through an experimental AI-enhanced writing studio, myProse, which has been developed based on research into the cognitive processes of writing and established principles of writing pedagogy.

myProse is specifically designed to support writers across the entire writing process, from initial drafting through multiple stages of revision and refinement. The platform fosters efficient drafting while carefully preserving the writer's agency. It offers sophisticated, tailored feedback on key aspects of writing such as argumentation strength, organizational coherence, and source credibility, delivered through an interactive visual-verbal interface that promotes active engagement with the revision process. Through natural language processing and machine learning algorithms, myProse analyzes textual patterns and provides targeted suggestions for improvement while maintaining the writer's authorial control. The design of myProse emphasizes the role of AI not as a replacement for human creativity but as a catalyst for productivity and innovation in writing, enabling writers to achieve their creative vision more effectively.

This talk advocates for a carefully balanced approach that embraces AI's potential as a sophisticated tool to empower writers while simultaneously safeguarding the indispensable contributions of human ingenuity to the writing process. Through this framework, we can harness AI's capabilities to enhance rather than diminish the fundamental human elements of writing, creating a partnership that advances both efficiency and creativity in written expression. The insights shared will contribute to ongoing discussions about responsible AI integration in writing education and practice.

**Keywords**: Writing Instruction, Generative AI, Writing Process

**References**

Kaufer, D., & Ishizaki, S. (2024). Future of Writing in the Disciplines and Professions. White Paper. Retrieved on December 23, 2024 from https://www.cmu.edu/dietrich/english/research-and-publications/myprose.html

Flower, L., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), Cognitive Processes in Writing (pp. 31–50). Lawrence Erlbaum Associates.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. College Composition and Communication, 32 (4), 365–38

**MARIO BARCALA** es Ingeniero Informático (1999) y Doctor en Computación (2010) por la Universidade da Coruña. Lleva más de 20 años trabajando e investigando, en colaboración con diferentes equipos de lingüistas, en diferentes temas relacionados con la lingüística de corpus y la lingüística computacional. Emplea la tecnología para solucionar problemas propuestos por los lingüistas y, desde el año 2015, realiza esta tarea desde su empresa, NLPgo Technologies, S.L.

También es formador e imparte diferentes talleres para lingüistas sobre temas relacionados con la informática, la lingüística computacional y la inteligencia artificial. La mayoría de estas iniciativas de formación la canaliza a través del blog "Palabras binarias: Informática para lingüistas", aunque también imparte talleres en otras entidades y Universidades.

**MARIO BARCALA** is a Computer Engineer (1999) and holds a PhD in Computing (2010) from Universidade da Coruña (Spain). He has been working and conducting research for over 20 years, collaborating with various teams of linguists on different topics related to corpus linguistics and computational linguistics. He is devoted to using technology to solve problems proposed by linguists, and since 2015, he has been carrying out this work at his company NLPgo Technologies, S.L.

He is also a trainer and delivers workshops for linguists on topics related to computer science, computational linguistics, and artificial intelligence. Most of these training initiatives are channeled through the blog "Palabras Binarias: Informática para Lingüistas" (Binary Words: Computer Science for Linguists), although he also conducts workshops for other organizations and universities.

## Tareas tradicionales, soluciones modernas: IA en la lingüística de corpus - PaCor 2025

La lingüística computacional ha afrontado desde sus inicios retos significativos en tareas fundamentales relacionadas con el procesamiento del lenguaje natural, como la transcripción, el reconocimiento de escritura (OCR) y la etiquetación morfosintáctica (POS *tagging*), entre otras. El enfoque clásico para solucionar estos retos ha estado tradicionalmente ligado al uso de técnicas basadas, principalmente, en modelos matemáticos y estadísticos (Rabiner et al., 1989) (Molinero et al., 2007), que se han combinado de diversas formas (Mohri et al., 2002) para intentar resolver diferentes tareas de la mejor forma posible.

Sin embargo, algunos problemas han estado permanentemente presentes en muchas de estas soluciones: dificultad del tratamiento de la variabilidad en la entrada (diferentes timbres de voz, ruido, diferentes modos de escritura, etc.) (García & Tapias, 2000), complejidad en la integración de las diferentes capas de procesado (modelo fonético/modelo lingüístico, etiquetación/lematización) (Rufiner & Milone, 2004) y eficacia limitada a idiomas y/o dominios de aplicación específicos.

Por otra parte, aunque las soluciones basadas en redes neuronales se han venido utilizando desde hace ya varios años (Mikolov et al., 2011), marcando un importante avance en diversas áreas, no ha sido hasta el reciente auge de la inteligencia artificial cuando se ha transformado significativamente el panorama tecnológico, principalmente debido al desarrollo de los *Transformers* (Vaswani, et al., 2017), que han impulsado la aparición de los modelos de lenguaje de gran tamaño. Estas innovaciones han permitido superar muchas de las limitaciones de las tecnologías disponibles hasta el momento (Humphries, 2024) (Chen, 2024) y lograr, no solo una notable mejora en la eficacia de las nuevas soluciones, sino también simplificar el acceso a estas herramientas.

En este trabajo compartimos nuestra experiencia en la prueba y el desarrollo de soluciones basadas en IA para resolver diferentes tareas vinculadas con la lingüística computacional, el procesamiento del lenguaje natural y la lingüística de corpus, exponiendo algunos casos prácticos relacionados con la transcripción, la asignación de interlocutores y la etiquetación morfosintáctica:

- **Transcripción**: La transcripción de audio a texto sigue siendo una tarea esencial para la creación de corpus orales y el análisis de fenómenos del habla. La transcripción manual de audio requiere un consumo de recursos humanos y de tiempo nada desdeñables, por lo que cualquier mejora en el terreno de la realización de la tarea de manera automática, que permita reducir el tiempo de revisión posterior, resulta crucial para reducir los costes y acelerar el avance de proyectos. Veremos que, empleando técnicas basadas en IA, hoy es posible usar modelos de reconocimiento automático de voz que alcanzan niveles de precisión impensables hace muy poco tiempo.
- **Asignación de interlocutores**: Las interacciones a través de redes sociales y aplicaciones de mensajería han abierto nuevas oportunidades para la investigación lingüística. Sin embargo, trabajar con estas fuentes plantea retos únicos. Un ejemplo de lo que ha permitido la IA en este campo es el de la creación de flujos de trabajo sencillos y prácticos para la construcción de corpus de conversaciones, como comprobaremos viendo un ejemplo de aplicación de recogida de interacciones procedentes de diferentes aplicaciones de mensajería y redes sociales.
- **Etiquetación morfosintáctica**: La etiquetación morfosintáctica es una de las tareas fundamentales presente en muchas aplicaciones relacionadas con el procesamiento del lenguaje natural. En este apartado intentaremos dar respuesta a las siguientes preguntas: ¿Es posible utilizar técnicas basadas en IA para realizar esta tarea? ¿Mejoran a las ya existentes? ¿Sigue siendo en la actualidad una etapa fundamental del procesamiento del lenguaje natural?

Los ejemplos mostrados en este trabajo no solo ponen de relieve los avances que pueden proporcionar las soluciones basadas en IA en algunos contextos, sino que también apuntan a la necesidad de realizar un cambio paradigmático en la manera en que abordamos algunos problemas. Si bien es cierto que las soluciones basadas en IA no siempre serán la solución más adecuada para todos los problemas, se han abierto nuevas posibilidades para proyectos que antes se consideraban inalcanzables, democratizando el acceso a herramientas avanzadas y ampliando el impacto potencial de la investigación lingüística.

El objetivo último de este trabajo es el de reflexionar, por una parte, sobre el alcance y las limitaciones de las tecnologías tradicionales frente a las basadas en IA en tareas clásicas relacionadas con la lingüística computacional, el procesamiento del lenguaje natural y la lingüística de corpus, y por otra, sobre la necesidad de cambiar el enfoque a la hora de resolver algunas tareas clásicas relacionadas con estas áreas de conocimiento.

**Palabras clave:** Inteligencia artificial, transcripción, asignación de hablantes, etiquetación morfosintáctica

**Referencias**

Chen C. et al. (2024) HyPoradise: An Open Baseline for Generative Speech Recognition with Large Language Models. arXiv:2309.15701, https://doi.org/10.48550/arXiv.2309.15701

García, C. & Tapias, D. La frecuencia fundamental de la voz y sus efectos en el reconocimiento de habla continua. (2000). *Procesamiento del lenguaje natural*, 26, 163-168.

Humphries, M. (2024). Unlocking the Archives: Large Language Models Achieve State-of-the-Art Performance on the Transcription of Handwritten Historical Documents. arXiv:2411.03340, https://doi.org/10.48550/arXiv.2411.03340

Mikolov, T et al. (2011). Strategies for training large scale neural network language models. *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2011, 196-201.

Mohri, M. et al. (2002). Weighted Finite-State Transducers in Speech Recognition. Computer Speech & Language, volume 16, issue 1, 69-88.

Molinero, M.A. (2007). Practical application of one-pass Viterbi algorithm in tokenization and part-of-speech tagging. *Recent Advances in Natural Language Processing*, 35-40.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, volume 27, issue 2, 257-286.

Rufiner, H.L. & Milone, D.H. (2004). Sistema de reconocimiento automático del habla. *Ciencia, Docencia y Tecnología*, vol. XV, núm 28, 151-177.

Vaswani, A. et al. Attention Is All You Need. (2017). *Advances in Neural Information Processing System*, 30.

**MARIE-AUDE LEFER** es Profesora Titular en Traducción y Traducción Inglés-Francés en UCLouvain, Bélgica, donde es decana de la Faculatd der Traducción er Interpretación de Lovaina. Sus actuales líneas de investigación incluyen nuevas tecnologías en la formación de traductores, entrenamiento en pos-edición de traducciones automáticas, estudios basados en corpus de traducciones de estudiantes y pos-edición, evaluación de calidad de pos-edición y traducción, marcación de errores de traducción, métodos de tarificación de pos-edición y remuneración justa.

Ha co-editado nueve volúmenes y números especiales, entre los que destacan: "Empirical Translation Studies: New Methodological and Theoretical Traditions" (De Gruyter, 2017), "Extending the Scope of Corpus-Based Translation Studies" (Bloomsbury, 2022) y "Learner Translation Corpus Research" (Benjamins, 2023).

**MARIE-AUDE LEFER** is Associate Professor of Translation Studies and English-French translation at UCLouvain, Belgium, where she acts as Head of the Louvain School of Translation and Interpreting. Her current research interests include technology in translator education, machine translation post-editing training, corpus approaches to student translation and post-editing, post-editing and translation quality assessment, translation error annotation, post-editing pricing methods, and fair pay.

She has co-edited nine volumes and special issues, such as *Empirical Translation Studies: New methodological and theoretical traditions* (De Gruyter, 2017), *Extending the Scope of Corpus-based Translation Studies* (Bloomsbury, 2022) and *Learner Translation Corpus Research* (Benjamins, 2023).

Sus publicaciones más recientes incluyen "The Machine Translation Post Editing Annotation System (MTPEAS): A Standarized and User Friendly Taxonomy for Student Post Editing Quality Assessment" (Translation Spaces) e "Introducing MTPE Pricing in Translator Training: A Concrete Proposal for MT Instructors" (The Interpreter and Translator Trainer).

Es co-directora del Proyecto de Multilingual Student Translator (MUST), una iniciativa internacional que involucra a 40 equipos de todo el mundo. MUST se centra en construir un amplio corpus multilingüe de traducciones de estudiantes, y actualmente incluye más de 8.000 traducciones. Como parte del proyecto MUST, co-desarrolló el Sistema de Anotación Orientado a la Traducción (TAS), diseñado para anotar errores en traducciones de estudiantes. Más recientemente, ha liderado el proyecto postedit.me, que busca recopilar y anotar textos pos-editados por estudiantes. Este proyecto ha generado dos recursos clave para la comunidad docente e investigadora: una aplicación para recopilar, anotar y buscar corpus de pos-edición de estudiantes, y el Sistema de Anotación de Pos-edición de Traducción Automática (MTPEAS) para evaluar la calidad de la pos-edición.

Her most recent journal publications include *The Machine Translation Post-Editing Annotation System (MTPEAS): A standardized and user-friendly taxonomy for student post-editing quality assessment* (Translation Spaces) and *Introducing MTPE Pricing in Translator Training: A Concrete Proposal for MT Instructors* (The Interpreter and Translator Trainer). She is the co-director of the *Multilingual Student Translation* (MUST) project, an international initiative involving 40 teams from around the world. MUST focuses on building a large multilingual corpus of student translations, which currently comprises over 8,000 translations. As part of the MUST project, she co-developed the Translation-oriented Annotation System (TAS), designed for error-annotating student translations. More recently, she has led the postedit.me project, which aims to collect and annotate student post-edited texts. This project has yielded two key resources for the research and teaching community: an app for collecting, annotating, and searching student post-editing corpora, and the Machine Translation Post-Editing Annotation System (MTPEAS) for assessing post-editing quality.

## Parallel corpora in the age of AI: Designing, collecting, annotating and analyzing machine translation post-editing corpora

To date, parallel corpora used in contrastive linguistics and translation studies have typically consisted of professional translations into the translators' native language, limited to a narrow range of text types, such as fictional prose and news (Lefer, 2020). These parallel corpora are often only superficially documented in terms of metadata, providing little information about translator profiles, communication contexts, or translation workflows (including the use of technology). In recent years, machine translation (MT) has seen a steady rise in the language service industry, further driven by the adoption of Large Language Models and generative AI for translation tasks. Simultaneously, machine translation post-editing (MTPE) has gained traction as a key language service provided by professional linguists. Consequently, translation "from scratch" (i.e. without relying on MT technology) can no longer be considered the default translation method. According to the European Language Industry Survey 2024, "it is expected that some form of MT or AI will be used in more than 50% of professional translations by 2025" (ELIS Research, 2024: 4). For corpus-based cross-linguistic research to remain relevant in today's context, it is essential to examine new translation practices shaped by technological advances, with MTPE serving as a prominent example. Scholars need to collect MTPE corpora to explore the key linguistic features of this form of translation and, more broadly, to investigate the principles underlying post-editing behavior as distinct from translational behavior.

In this talk, I will present the Post-Edit Me! project (Lefer et al., 2023), which focuses on collecting, annotating, and analyzing a large parallel corpus of student post-edited texts. First, I will describe the main features of the postedit.me app, which we developed to facilitate the collection, annotation, alignment, and querying of MTPE texts (Lefer et al., 2024). Next, I will outline the Machine Translation Post-Editing Annotation System (MTPEAS), which we designed to systematize the analysis of MTPE edits – i.e. changes made to raw MT output – and to standardize the evaluation of MTPE quality (Lefer et al., 2022; Bodart et al., 2024). MTPEAS incorporates a taxonomy of seven categories that encompass typical post-editing edits: Value-adding edits, Successful edits, Unnecessary edits, Incomplete edits, Error-introducing edits,

Unsuccessful edits, and Missing edits. Both resources – the app and the MTPEAS annotation manual – are freely available to the research community for internal research and teaching purposes. Finally, I will present the initial findings from analyses of the Post-Edit Me! corpus for the English-French language pair, comparing translations and post-edited texts produced by translation students, with a particular focus on translation errors. I will conclude by taking stock of the project to outline the main theoretical, descriptive, and applied insights that can be derived from the analysis of (student) MTPE corpora.

Taller / Workshop

**JOSÉ MANUEL FRADEJAS RUEDA**, obtuvo la licenciatura en Lingüística Hispánica en la Universidad Complutense de Madrid en 1980 y se doctoró en 1983 con una tesis titulada *Tratado de cetrería. Texto, gramática y vocabulario*, dirigida por el académico Manuel Alvar. Su carrera docente comenzó en la Universidad Nacional de Educación a Distancia (UNED), donde fue profesor titular de Lengua Española y Filología Románica entre 1981 y 2000. Desde 2009 es catedrático en la Universidad de Valladolid.

Su interés por la cetrería lo llevó a fundar y dirigir el Archivo Iberoamericano de Cetrería, una plataforma digital que recopila y difunde fuentes sobre esta práctica en el mundo iberorrománico .

En el ámbito de la Filología Digital, ha sido pionero en la aplicación de tecnologías como la codificación XML/TEI para la edición de textos medievales.

**JOSÉ MANUEL FRADEJAS RUEDA** earned his degree in Hispanic Linguistics from the Universidad Complutense of Madrid in 1980 and completed his PhD in 1983 with a dissertation entitled *Tratado de cetrería. Texto, gramática y vocabulario*, supervised by the scholar Manuel Alvar. His academic career began at the *Universidad Nacional de Educación a Distancia* (UNED), where he served as Associate Professor of Spanish Language and Romance Philology from 1981 to 2000. Since 2009, he has been a full professor at the Universidad de Valladolid.

His interest in falconry led him to establish and lead the *Archivo Iberoamericano de Cetrería*, a digital platform that collects and disseminates sources on this practice throughout the Ibero-Romance milieu.

In the field of Digital Philology, he has been a pioneer in the application of technologies such as XML/TEI encoding for the edition of medieval texts.

Destaca su proyecto 7 Partidas Digital, que tiene como objetivo la edición crítica digital de las Siete Partidas de Alfonso X, incorporando análisis estilométricos y herramientas de procesamiento de lenguaje natural.

José Manuel Fradejas Rueda ha impartido numerosos cursos y conferencias sobre la codificación de textos con TEI, estilometría computacional y análisis textual utilizando el lenguaje de programación R.

Fruto de su experiencia docente y su compromiso con la innovación en las Humanidades Digitales, ha publicado el libro *Cuentapalabras: Estilometría y análisis de texto con R para filólogos*. Esta obra, disponible en línea, ofrece una introducción práctica al análisis automatizado de textos con R, dirigida especialmente a filólogos y humanistas interesados en aplicar técnicas cuantitativas a la investigación textual .

One of his most notable projects is 7 Partidas Digital, which aims to produce a digital critical edition of Alfonso X's *Siete Partidas*, integrating stylometric analysis and natural language processing tools.

José Manuel Fradejas Rueda has delivered numerous courses and lectures on TEI text encoding, computational stylometry, and textual analysis using the R programming language.

As a result of his teaching experience and his commitment to innovation in the Digital Humanities, he has published the book *Cuentapalabras: Estilometría y análisis de texto con R para filólogos*. This open-access publication offers a hands-on introduction to automated text analysis with R, specifically designed for philologists and humanists interested in applying quantitative methods to textual research.

## Introducción a R para el análisis lingüístico de textos: una aproximación desde los corpus de discursos políticos

Este taller introductorio está dirigido a investigadores, estudiantes y profesionales del ámbito de la lingüística del corpus interesados en aprender a utilizar R como herramienta para la extracción y el análisis de información lingüística a partir de textos. El objetivo es proporcionar una base práctica y accesible para comenzar a trabajar con datos textuales en R, sin necesidad de conocimientos previos en programación.

A lo largo del taller, se trabajará con una serie histórica de discursos políticos pronunciados por los presidentes de los Estados Unidos desde 1790 hasta la actualidad (2025). Esta colección textual permite observar, desde una perspectiva sincrónica y diacrónica, fenómenos lingüísticos como el uso de ciertas construcciones gramaticales, el léxico dominante en cada periodo, la evolución de los recursos retóricos y estilísticos, o la presencia de determinadas temáticas. Se mostrará también cómo obtener metadatos relevantes (año, presidente, contexto histórico) que permiten enriquecer el análisis.

Durante el taller se presentarán los conceptos básicos del entorno de R y se introducirán herramientas específicas para el análisis de corpus, como los paquetes *tidytext, quanteda, y stringr*, entre otros. Se explicará cómo importar y estructurar los textos, realizar búsquedas léxicas, generar estadísticas descriptivas, visualizar frecuencias o nubes de palabras, y representar gráficamente ciertos patrones lingüísticos.

Aunque el corpus principal será el de los discursos presidenciales estadounidenses, se discutirán brevemente otras posibilidades, como los discursos de Navidad del Rey de España, que pueden servir para ejercicios alternativos, aunque su extensión temporal es más limitada.

El enfoque del taller será eminentemente práctico: los participantes podrán seguir los ejemplos en sus propios ordenadores y reproducir los análisis con los materiales proporcionados. Se entregarán scripts comentados y recursos para profundizar tras la sesión.

Este taller no solo busca introducir a los participantes en el uso de R, sino también fomentar una reflexión crítica sobre el potencial de las herramientas cuantitativas y computacionales en el estudio del lenguaje. A través del análisis de discursos institucionales, se ofrece un marco aplicable a otros corpus y contextos lingüísticos.

# 1. Aplicaciones de la IA generativa en la traducción y los corpus multilingües / Generative AI applications in translation and multilingual corpora

# Estudio sobre la aceptabilidad de la traducción de metáforas de animales en el discurso político chino: una comparación entre la traducción oficial y ChatGPT

*FU, Xiaoqiang (Guangzhou Xinhua University, China)*

Como elemento esencial del lenguaje y la cultura, las metáforas no solo constituyen un recurso retórico, sino que también transmiten significados culturales y sociales profundos (Lakoff y Johnson, 1980; Kövecses, 2010). En el discurso político chino, las metáforas han adquirido un papel destacado como recurso estilístico recurrente. Entre ellas, las metáforas de animales se distinguen por su carácter vívido y accesible, lo que les confieren una notable efectividad en la comunicación de mensajes políticos. Sin embargo, debido a su fuerte arraigo en contextos culturales específicos, la traducción de metáforas, particularmente en textos políticos, representa un desafío significativo en el ámbito de la traducción interlingüística (Newmark, 2010; Chico Rico, 2015).

En los últimos años, con el rápido desarrollo de la tecnología de inteligencia artificial, las herramientas de traducción automática han adquirido un papel cada vez más destacado en la práctica de la traducción. En particular, los traductores basados en inteligencia artificial generativa, representados por ChatGPT, han despertado un amplio interés y aplicación tanto en el ámbito académico como en el profesional, debido a su desempeño en la traducción de múltiples lenguas y disciplinas (Geng y Hu, 2023; Jin, 2024). Sin embargo, en comparación con las traducciones oficiales realizadas por el Instituto de Historia y Literatura del Partido del Comité Central de China, aún falta una investigación empírica que evalúe la capacidad de estas herramientas para manejar metáforas en textos políticos, así como su grado de comprensión en la lengua meta y su naturali ultural en este ámbito.

**Palabras clave:** traducción de ChatGPT; metáforas de animales; traducción chino-español; traducción de metáforas

**Referencias en español:**

Calvo-Ferrer, J. R. (2023). Can you tell the difference? A study of human vs machine-translated subtitles. Perspectives, 32(6), 1115–1132.

Chico Rico, F. (2015). La traducción del texto político: características pragmático-discursivas y estrategias traductológicas. *Tonos Digital*, 29(0), 1-25. http://www.tonosdigital.es/ojs/index.php/tonos/article/viewFile/1308/784

Gao, R. *et al* (2024). Machine translation of Chinese classical poetry: a comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications*, 11(1), 1-10.

Jiang, L., Jiang, Y., & Han, L. (2024). The potential of ChatGPT in translation evaluation: A case study of the Chinese-Portuguese machine translation. *Cadernos De Tradução*, 44(1), 1-22.

Kövecses, Z. (2010). *Metaphor. A Practical Introduction* (2ª ed). New York: Oxford University Press.

López Luis, M. (2022). *La traducción de discursos políticos en contextos institucionales. El análisis del discurso aplicado a la traducción inglés/español de Barac H. Obama y Donald J. Trump*. Tesis doctoral, Universidad de Córdoba.

Lakoff, G. y J. Mark. (1980) . *Metaphors We Live By*. Chicago: University of Chicago Pres.

Luque Janodet, F. (2020). La metáfora conceptual en el discurso político euroescéptico (francés-español). *Logos: Revista de Lingüística, Filosofía y Literatura*, 30(2), 349-364. http://dx.doi.org/10.15443/rl3026

Mendoza García, I. (2015). La aceptabilidad de la traducción cultural en la literatura para la infancia: una propuesta conceptual y metodológica. *Tonos digital: revista de estudios filológicos*, 29, 1-25.

Newmark, P. (2010). *Manual de traducción* (6ª ed.), (Traducido por Virgilio Moya). Madrid: Catedra.

Pragglejaz, G. (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. Metaphor and Symbol, 22(01), 1-39.

Spoturno, M. L. (2024). Traducción literaria e inteligencia artificial: consideraciones para la formación universitaria. Cadernos de Tradução. 44, 1-26. https://doi.org/10.5007/2175-7968.2024.e100602.

**Referencias en chino:**

Bai, Y. y Zhang, W. (白一博，张威) (2024). El contenido narrativo y las estrategias de traducción de la metáfora de "guerra" en el discurso

político: Un estudio basado en La gobernación y administración de China (Volumen IV) (战争隐喻叙事的内涵与翻译策略——以《习近平谈治国理政》（第四卷）为例). *Revista china de ESP* (《中国ESP研究》), 36, 1-12.

Chen, S. (陈双双) (2019). Estudio sobre las características de la traducción de documentos oficiales chinos al exterior (《中央文件对外翻译的特点研究》). *Revista de la Universidad de Zhongzhou* (《中州大学学报》), 35(3), 69-74.

Chen, X. (陈小慰) (2016). La "perspectiva retórica" en la traducción de documentos centrales (《中央文献翻译中的"修辞观"》). *Estudio de traducción de chino* (《中译外研究》), 5, 29-40.

Dou, W. y Du, H. (窦卫霖，杜海紫) (2018). Estudio sobre la aceptabilidad de la traducción de nuevos términos populares en la China contemporánea (《中国当下流行新词翻译的可接受性研究》). *Revista de la Universidad de Este China* (《华东师范大学学报》), 6, 65-71.

Geng, F. y Hu, J. (耿芳，胡健) (2023). Nuevas direcciones en la posedición asistida por inteligencia artificial: Un estudio de casos de traducción basado en ChatGPT (《人工智能辅助译后编辑新方向——基于ChatGPT的翻译实例研究》). *Legua extranjera de China* (《中国外语》), 111(3), 41-47.

Jin, Y. (金艳玲) (2024). Estudio comparativo de traducción de textos históricos chinos mediante traducción automática y generación de IA: Papago y ChatGPT como ejemplos (《机器翻译与生成式AI汉朝翻译对比研究——以Papago翻译器与ChatGPT为例》). *Korean Language in China*, 6, 76-85.

Liu, L. y Zhang, S. (刘立新，郑淑明) (2024). Estudio de la traducción de cadenas de cohesión metafórica en discursos políticos: Un análisis de las traducciones al inglés de La gobernación y administración de China (Volúmenes I y II) (《政治语篇隐喻衔接链条的翻译研究——以〈习近平谈治国理政〉（第1、2卷）英译为例》). *Traducción y comunicación* (《翻译与传播》), 1(9), 97-112.

Ren, D. (任东升) (2016). Exploración preliminar sobre la escritura histórica de la práctica de la traducción estatal: Una revisión de los "traductores extranjeros" en la práctica de la traducción estatal (《国家翻译实践史书写的初步探索——国家翻译实践中的"外来译家"研究综述》). *Investigación*

*sobre la escritura de la historia de la práctica de China* (《国家翻译实践史书写研究》), 5, 1-5.

Xiong, D. (熊道宏) (2019). Preguntas y traducción: Reflexiones sobre el lenguaje político y los mecanismos de trabajo (《答疑与翻译——对政治语言与工作机制的思考》). Revista de la Universidad de Estudios Internacionales de Tianjin (《天津外国语大学学报》), 25(2), 43-52.

Xi, Jinping (习近平) (2014-2020). *La gobernación y administración de China* (Volúmenes I-III) (《习近平谈治国理政》). Beijing: Waiwen Chubanshe.

Xi, Jinping (习近平) (2014-2021). *La gobernación y administración de China* (Volúmenes I-III, edición en español). Beijing: Waiwen Chubanshe.

Wen, X. y Tian, Y. (文旭，田亚灵) (2024). Estudio sobre la efectividad de la aplicación de ChatGPT en la traducción de discursos con características chinas (《ChatGPT应用于中国特色话语翻译有效性研究》). *Revista de traductología de Shanghai* (《上海翻译》), 2, 27-34.

Zhu, W. y He, N. (朱炜，贺宁杉) (2011). Metáfora y construcción de discursos políticos (《隐喻与政治语篇的建构》). *Revista de la Universidad de Nanchang* (《南昌大学学报》), 42(3), 121-125.

# 2. Codificación de corpus / Corpus encoding

# Presentation of MedCor, an aligned parallel corpus (English-Spanish) of medical fictional language: compilation, annotation and possible applications

*Goretti Faya Ornia (Universidad de Valladolid)*

*Patricia Rodríguez Inés (Universitat Autònoma de Barcelona)*

MedCor is an aligned parallel corpus of subtitles from medical TV programmes (Dr House, Grey's Anatomy and ER) in English and Spanish, built by the Research Group "Communicative and Intercultural Skills in Foreign Language" at the University of Valladolid. The strings were manually checked and aligned. At this point, following the annotation recommendations of Zanettin (2012), a two-step annotation is being carried out. First, an automated POS annotation was performed in both languages. Secondly, a documental annotation will be carried out, including some metadata and taking into account not only the bibliographic information of each episode, but also the type of interlocutors (i.e. doctor-doctor or doctor-patient), as this can have an important impact on the tone and vocabulary used and it can become a useful tool for discursive studies. Also, a web searcher is being developed to search the data in a more user-friendly way and with more options than traditional databases.

Aligned parallel corpora are used in a wide range of fields and disciplines, and have diverse purposes (Gallego Hernández, 2011; Leiva Rojo, 2018). In this sense, MedCor has many potential applications and users, as it could be useful for academics (either in linguistics or translation studies), for teachers and students (to learn a foreign language or some translation strategies), and for professional translators and software developers (as a bilingual resource to expand their databases). So far, the corpus has been used as a research tool and a didactic resource in the teaching of (a) English as a foreign language, (b) medical English and (c) translation. Focus groups formed by students were organised in the classroom to assess the didactic impact of the corpus, and their insights were recorded in a systematic grid. So far the results have been promising, with students finding the corpus an interesting and motivating learning resource. The main aim of this presentation is therefore to introduce MedCor, explain its compilation and annotation process, and discuss its various applications, purposes and preliminary results.

**Keywords:** Parallel corpus, compilation, annotation, medical translation, audiovisual, subtitles.

## References

Gallego-Hernández, Daniel (2018). New Insights into Corpora and Translation. Cambridge Scholars Publishing.

Leiva Rojo, Jorge Jesús (2018). Diseño y compilación de corpus paralelos alineados: dificultades y (algunas) soluciones en el ejemplo de un corpus de textos museísticos traducidos (inglés-español). *Revista de lingüística y lenguas aplicadas*, 13, 59-73.

Zanettin, Federico (2012). Translation-driven corpora: corpus resources for descriptive and applied translation studies. Routledge.

# Bridging Corpus Linguistics and Data Science: A Multifactorial Study of English-to-Russian Translation of Reporting Verbs in Literary Texts

*Łukasz Grabowski (University of Opole, Poland)*

*Daniel Borysowski (University of Opole, Poland)*

Nowadays in many corpus studies researchers make ample use of machine learning techniques popular in the field of natural language processing (NLP), also known as computational linguistics, as well as data science, which has become, in a sense, a separate discipline devoted to data analysis. This allows us to put forward not only descriptive or interpretative hypotheses, but also explanatory or predictive ones (e.g. Gries & Wulff 2012; van Beveren & De Sutter 2019; Oakes 2019; De Sutter & Lefer 2019; De Sutter et al. 2023; Chmiel & Kajzer-Wietrzny 2024). For example, without the use of different regression methods (depending on the type of a dependent variable), it would have been impossible to verify the hypotheses explaining the occurrence of some linguistic features with a reliable degree of precision.

Examining repeated patterns of language use is a cornerstone of much corpus linguistic research, also oriented at translation, where repetition – manifested on several linguistic levels (morphological, syntactic, lexical etc.) – plays a very important role, notably in literary texts. In this study, our goal is to identify the predictors of repetition or lexical variety in the translation of reporting verbs from English into Russian. The main reason for selecting reporting verbs in this study was that the authors of literary novels use a wide array of reporting verbs following dialogues (e.g. *said, muttered, murmured, whispered, sighed, gasped*), which - due to their high frequency in such texts - are particularly well-suited to wide-range statistical analyses (cf. Mastropierro 2020, 2022 for research in the English-Italian language pair). Using a sample of 20 literary novels, we fit multiple negative binomial regression with mixed effects to assess the effect that selected predictor variables (e.g. frequency of a source-text verb, its number of senses in Princeton WordNet (domain: communication), semantic type of the verb, length of the verb in characters, date of translation, and translators) have on the response variable: the number of different target language verb types (lemmas) a source text reporting verb is translated into Russian. Intuitively, if the number of types is way higher than 1 then it means that translators opted for lexical variety (i.e. used various TT reporting verbs as translation equivalents of a single repeatedly used ST verb).

For instance, the prosodic (cf. Caldas-Coulthard (1987) typology) verb *shouted* was translated into 4 different reporting verbs in Russian (***закричал***, ***воскликнул***, ***кричал***, ***крикнул***) in the novel *The Fellowship of the Ring* when signalling direct speech following dialogues, which indicates lexical variety. Reporting verbs were retrieved using custom-designed CQL queries from InterCorp v.15 (Čermák & Rosen 2012; Čermák 2019), a large multilingual parallel corpus which includes, among others, English novels and their translations into Russian. This way, we also complement the findings of earlier research conducted in English-to-Italian (Mastropierro 2020; 2022) and English-to-Polish language pair (Mastropierro & Grabowski 2024)

Preliminary findings: The overall model fit per the lowest AIC and BIC values obtained through backward elimination reveals that semantic category of a ST reporting verb, its frequency and translation date as well as the translator as a random effect have the largest individual contributions to explaining the proportion of variation in the response variable in the Russian translations. More precisely, the model allows us to explain almost 73% of variation (conditional r-squared = 0.73) in the response variable, that is, the number of different verb types a ST verb is translated into. We also identified semantic types of English ST verbs (i.e. structuring verbs, e.g. *ask* or *reply*, and signalling discourse verbs, e.g. *repeat* or *add*; cf. the typology by Caldas-Coulthard 1987) whose Russian equivalents tend to be consistently repeated in translation. In other words, such verbs were consistently rendered with a narrow range of the same Russian equivalents, which means more repetition in translation. The low level of variance (0.05) in the random effect means that the impact of individual translators is relatively similar. In other words, there is some variability between the translators, but it is relatively small, and no single translator significantly influenced overall results.

We conclude with discussion of our preliminary findings and of limitations of the study itself. We also outline avenues for future research, as the methodology could be easily adopted to other language pairs, including English-to-Spanish or Spanish-to-English translation. Overall, we hope that our study, notably its novel methodology, will inspire further research on predictors of translatorial decisions, also including factors related to technological developments, including AI-assisted translation, that have recently become an integral element also of the translator's toolkit and as such may impact translation style in the near future.

**Keywords:** corpus linguistics, literary translation, parallel corpus, mixed-effects, regression modelling

**Acknowledgements:** This study is funded by the National Science Centre (NCN), Poland, grant number: 2023/51/B/HS2/00697.

## References

Caldas-Coulthard, Carmen R. 1987. Reported speech in written narrative texts. In Malcolm Coulthard ed. *Discussing Discourse*. Birmingham: University of Birmingham, 149–167.

Chmiel, Agnieszka and Marta Kajzer-Wietrzny. 2023. Into B or not into B? The limited impact of interpreting direction on target text fluency and complexity. *SKASE Journal of Translation and Interpretation* 16(2): 23–43.

De Sutter, Gert and Marie-Aude Lefer. 2019. On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives* 28(1): 1–23.

De Sutter, Gert, Marie-Aude Lefer and Bram Vanroy. 2023. Is linguistic decision-making constrained by the same cognitive factors in student and in professional translation? Evidence from subject placement in French-to-Dutch news translation. *International Journal of Learner Corpus Research* 9(1): 60–95.

Gries, Stefan and Stefanie Wulff. 2012. Regression analysis in translation studies. In: M. Oakes & M. Li eds. Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research. Amsterdam: John Benjamins, 35–52.

Mastropierro, L. (2020). The translation of reporting verbs in Italian: The case of the Harry Potter series. *International Journal of Corpus Linguistics*, *25*(3): 241–269.

Mastropierro, L. (2022). The avoidance of repetition in translation: A multifactorial study of repeated reporting verbs in the Italian translation of the Harry Potter series. In L. Defang, & R. Moratto (Eds.), *Advances in corpus applications in literary and translation studies* (pp. 138–157). London: Routledge.

Mastropierro, L. & Grabowski, Ł. (2024). Repeated reporting verbs in English novels and their Italian and Polish translations: A preliminary multifactorial study. *Across Languages and Cultures*, 25(2): 310-330.

Oakes, Michael. 2019. Statistics for Corpus-Based and Corpus-Driven Approaches to Empirical Translation Studies. In Michael Oakes and Meng Li eds. *Advances in Empirical Translation Studies: Developing Translation*

*Resources and Technologies.* Cambridge: Cambridge University Press, 28–52.

Van Beveren, Amelie, Gert De Sutter and Timothy Colleman. 2019. The Mechanisms Behind Increased Explicitness in Translations: A Multifactorial Corpus Investigation of the Om-Alternation in Translated and Original Dutch: In Lore Vandervoorde, Joke Daems and Bart Defrancq eds. *New Empirical Perspectives on Translation and Interpreting*. New York: Routledge, 28–66.

# Potenciar la Anotación Pragmática con los Grandes Modelos de Lenguaje (LLM): Aceleración del Proceso y Contextualización con GPT-4

*Raffaella Gambardella (Università degli Studi di Salerno, Italy)*

En los últimos años, los Grandes Modelos de Lenguaje (LLM) han tenido un enorme éxito en diversos campos de investigación. La arquitectura básica de la mayoría de los Grandes Modelos de Lenguaje (LLM) se basa en los denominados tranformers (Vaswani et al., 2017). Aunque los detalles exactos de la arquitectura de GPT-4 aún no están bien delineados, sabemos que el modelo subyacente es un transformers entrenado capaz de predecir el siguiente token dentro de una secuencia (OpenAI, 2024).

Esta investigación específica pretende explorar las posibilidades que ofrece la inteligencia artificial y, gracias al apoyo de los Grandes Modelos de Lenguaje (LLM) y del modelo GPT-4, se está intentando proceder a la anotación pragmática de diálogos en lengua española (corpus DiEspa - Diálogos en Español) (ParlarItaliano, 2008). En la base del proyecto se encuentra el esquema de anotación de diálogos Pr.A.T.I.D. (Pragmatic Annotation Tool for Italian Dialogues) (Savy, 2010), que en el momento de su creación requería el uso de un software de anotación manual (XGate), ahora obsoleto. Dado que ya no es posible anotar con el software XGate (Cutugno, D'Anna, 2006), se decidió intentar entrenar el modelo GPT-4 con el objetivo de acelerar la anotación pragmática de diálogos y refinar su memoria. Tradicionalmente, la anotación pragmática requiere una intensa actividad humana para identificar y clasificar los actos lingüísticos y los fenómenos pragmáticos complejos.

Mediante un enfoque iterativo, el modelo se adaptó para reconocer y anotar fenómenos pragmáticos específicos basados en el esquema de anotación de diálogos Pr.A.T.I.D. La metodología adoptada implica el preprocesamiento de los diálogos para adaptarlos a los requisitos del esquema Pr.A.T.I.D. (Castagneto, 2012), seguido del uso del modelo GPT-4 para la identificación preliminar de los moves dialógicos (Savy, Solís García, 2009) esperados que deben detectarse en el texto del diálogo. Posteriormente, las anotaciones generadas automáticamente fueron revisadas y corregidas por expertos humanos, garantizando así un alto nivel de precisión y fiabilidad. El trabajo de revisión permitió no sólo identificar posibles errores, sino corregirlos con el objetivo de perfeccionar el sistema y permitir que el modelo GPT-4 mejorara su rendimiento.

Además, se está probando la anotación mediante la actualización 4.5 de OpenAI. Se trata del nivel más avanzado de IA hasta la fecha, conocido internamente con el nombre en clave de

«Orion», que según la empresa puede proporcionar una interacción más natural capaz de entender las indicaciones del usuario de una forma «más humana». El nuevo modelo también incluye inteligencia emocional mejorada (sentiment analysis), lo que lo hace útil para tareas como mejorar la escritura, programar y resolver problemas prácticos, así como una reducción significativa de las «alucinaciones» en comparación con los modelos anteriores. De hecho, en versiones anteriores era frecuente encontrar «alucinaciones» en la aplicación del modelo Pr.A.T.I.D. con la introducción totalmente inventada de moves no previstas por la norma.

En conclusión, este estudio demuestra el potencial de los Grandes Modelos de Lenguaje (LLM) y del modelo GPT-4 para superar algunas de las limitaciones asociadas a la anotación pragmática manual. El objetivo es allanar el camino para futuras investigaciones sobre la aplicación de modelos generativos del lenguaje en otras áreas de la lingüística computacional y la posibilidad de ampliar este enfoque a otras lenguas y tipos de texto. Las implicaciones de este estudio se extienden más allá del campo de la anotación pragmática, sugiriendo nuevas direcciones para la integración de los Grandes Modelos de Lenguaje (LLM) en los procesos de análisis lingüístico.

**Keywords:** grandes modelos de lenguaje, anotación pragmática, GPT-4, diálogos

**References**

Castagneto Marina, (2012), “Il sistema di annotazione Pra.Ti.D tra gli altri sistemi di annotazione pragmatica. Le ragioni di un nuovo schema”, in ANNALI del Dipartimento di Studi Letterari, Linguistici e Comparati Sezione Linguistica, Università degli Studi di Napoli “L’Orientale”, Napoli, Italia.

Cutugno Francesco, D'Anna Leandro, (2006) “XGate e XRG: strumenti per l'editing visuale, l'interrogazione e il benchmarking di annotazioni linguistiche XML”, Dipartimento di Fisica - Gruppo NLP, Università 'Federico II' di Napoli, Italia, Dipartimento di Linguistica e Letteratura, Università di Salerno, Italia.

OpenAI, (2024), Arxiv, URL: https://arxiv.org/abs/2303.08774

ParlarItaliano, Studium Dipsum, (2006), URL: https://parlaritaliano.studiumdipsum.it/it/792-corpus-diespa-dialogos-en-espanol

Savy Renata, (2010) "Pr.A.T.I.D.: a coding scheme for pragmatic annotation of dialogues", Dipartimento di Linguistica e Studi Letterari dell'Università di Salerno, Fisciano (Salerno), Italia.

Savy Renata, Solís García Inmaculada (2008), "Strategie pragmatiche in italiano e spagnolo a confronto: una prima analisi su corpus", Università degli Studi di Salerno, Fisciano (Salerno), Italia.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N.

Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need." In: Advances in Neural Information Processing Systems. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://arxiv.org/abs/1706.03762

# PRAGMACOR: Corpus Multimodal Anotado de Conversaciones Telefónicas con Intérprete

*Raquel Lázaro Gutiérrez (Universidad de Alcalá, Spain)*

Esta contribución ofrece resultados del proyecto PRAGMACOR, ejecutado por la Universidad de Alcalá y financiado por la Agencia Española de Investigación (Pragmática de corpus e interpretación telefónica: análisis de ataques contra la imagen, Ref. PID2021-127196NA-I00). Aplica una metodología basada en la pragmática de corpus para abordar el estudio de los ataques contra la imagen en la interpretación telefónica con el objetivo último de desarrollar materiales de formación y herramientas de interpretación asistida por ordenador para intérpretes telefónicos.

Las conversaciones telefónicas constituyen un ejemplo de comunicación a distancia en contraposición a la comunicación cara a cara. Además, las interacciones interpretadas pueden considerarse asíncronas, ya que el destinatario no recibe el discurso original, sino la traducción del mismo después de que un intérprete lo haya elaborado en el idioma requerido. Algunos autores ya han señalado una mayor de ataques contra la imagen (face-threatening acts, FTA) tanto en la comunicación a distancia como en la asíncrona (Castro Cruz, 2017; Locher, 2010; Simmons, 1994). Su alta prevalencia y las dificultades de los ataques contra la imagen para los intérpretes telefónicos ya han sido señaladas por Lázaro Gutiérrez y Cabrera Méndez (2018) y Lázaro Gutiérrez (2017).

En nuestro proyecto examinamos los ataques contra la imagen de un corpus de conversaciones mediadas por intérpretes telefónicos entre proveedores de servicios y usuarios finales siguiendo la clasificación de Brown & Lenvinson (1978/1987). La gran complejidad de los ataques contra la imagen ya se ha descrito como un reto a la hora de realizar investigaciones cuantitativas sobre ellos (Clancey y O'Keeffe 2019). Para abordar este desafío, proponemos una metodología de análisis de corpus, que fusiona, siguiendo a Haugh (2014) y Santamaría García (2011), la pragmática interaccional con la pragmática de corpus.

Un primer paso ha sido la compilación de un corpus multilingüe y multimodal (texto y audio) de conversaciones que incluye, junto con el español, algunas de las lenguas más frecuentes en las que se prestan servicios de interpretación telefónica, a saber, inglés, francés, alemán y chino (Lázaro Gutiérrez, 2021). Tras una larga e imprescindible fase de establecimiento de acuerdos y protocolos

de gestión de datos junto con empresas de servicios de interpretación en España, las conversaciones se han recopilado, se han anonimizado y transcrito de forma semiautomática, se han sincronizado con los audios correspondientes y se han anotado en EXMaRALDA (Schmidt, 2005). EXMaRALDA es un software de gestión de corpus que permite presentar las transcripciones en un formato similar al de una partitura musical, muestra eventos paralelos en distintos niveles y es de código abierto y multiplataforma. También permite definir metaetiquetas para cada transcripción. El trabajo de anotación se ha desarrollado durante dos años y, con esta contribución, pretendemos presentar su primera versión.

**Palabras clave:** corpus anotado multimodal, pragmática de corpus, EXMaRALDA, interpretación telefónica

**Referencias**

Brown, P., Levinson, S. (1987/1978) Politeness. Some Universals in Language Usage, Cambridge, Cambridge University Press.

Castro Cruz, M. (2017) Ataque a la imagen y descortesía en los comentarios de blogs en español peninsular. Philologia Hispalensis; 31(1):37-63.

Clancey, A., O'Keeffe, B. (2019) "Corpus pragmatics", in A. Clancey / B. O'Keeffe / S. Adolphs (eds) *Introducing Pragmatics in Use,* New York, Routledge, 47-68.

Haugh, M. (2014) "Jocular Mockery as Interactional Practice in Everyday Anglo-Australian Conversation". *Australian Journal of Linguistics*. 34: 1. 76-99.

Lázaro Gutiérrez, R. (2017) "El estudio de la cortesía en conversaciones mediadas por un intérprete sanitario" (in English, "Research about politeness in conversations mediated by a healthcare interpreter", in E. Ortega Arjonilla (ed.) *Sobre la práctica de la traducción y la interpretaciónen la actualidad* (Vol. 2.: T. Barceló Martínez / I. Delgado Pugés, I. (eds.) *De traducción jurídica y socioeconómica e interpretación para los servicios públicos*), Granada, Comares. 265-276.

Lázaro Gutiérrez, R. (2021) Analysis of Face-Threatening Acts against Telephone Interpreters. The Interpreters' Newsletter; 26.

Lázaro Gutiérrez, R., Cabrera Méndez, G. (2018) Pragmática e interpretación telefónica: un estudio sobre ataques contra la imagen de los intérpretes

(FTA, Face threatening acts). EPiC Series in Language and Linguistics; 3:85-90.

Locher, M. (2010) Introduction: Politeness and impoliteness in computer-mediated communication. J Politeness Res.; 6(1):1-5.

Santamaría-García, C. (2011) "Bricolage assembling: CL, CA and DA to explore the negotiation of agreement in English and Spanish conversation" in Farr, F. and O'Keeffe, A. (eds.) *Applying Corpus Linguistics. Special issue of International Journal of Corpus Linguistics*. 16: 3. 345-370.

Schmidt, T. (2005). Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. Arbeiten zur Mehrsprachigkeit, Folge B, 62.

Simmons, T. L. (1994) Politeness Theory in Computer Mediated Communication [Master's thesis]. Birmingham: Aston University.

# Evaluation of Automatic Sentence Alignment Methods for Spanish-English, Spanish-German, and Spanish-Chinese Literary Texts

*Irene Doval (Universidad de Santiago de Compostela, Spain)*

*Michael Lang (PaCorES Group, Universidad de Santiago de Compostela, Spain)*

The PaCorES collection comprises three parallel bilingual bidirectional corpora: Spanish/German, Spanish/English, and Spanish/Chinese (1). The core corpora of the collection consist of literary texts from the late 20th and early 21st centuries, which were manually selected and sentence-aligned with their corresponding translations.

The fundamental step in creating parallel corpora is the alignment. Sentence alignment is the issue of finding correspondence between source sentences and their equivalent translations in the target text.

Recent advances in automatic alignment tools, including neural network-based methods have achieved accuracy levels between 90% and 95% for closely related languages like German or English. However, these methods are primarily optimized for non-literary texts, and their accuracy declines significantly with literary texts, necessitating manual revision

The challenges of sentence alignment are especially pronounced in the Spanish/Chinese language pair due to significant structural and linguistic differences.

This paper describes methods aimed at minimizing the need for subsequent manual revision. To address frequent misalignments caused by improper segmentation, we developed a Python script (2) tailored to the specific linguistic characteristics of each language. We evaluated three well-known tools for sentence alignment: LF-Aligner (Hunalign), Vecalign, and Bertalign (3). Aligning bilingual literary poses unique challenges, since most of the translation is interpretative and not based on 1-to-1 mappings between source and target sentences. Existing alignment methods have difficulty coping with 1-to-many and many to-many alignments that are common in literary texts.

We evaluated the performance of each aligner using standard metrics: precision, recall, and F1 score (the harmonic mean of precision and recall). These metrics were calculated for each of the three language pairs.

**Keywords:** bilingual corpus ; literary corpus ; parallel corpus ; sentence alignment, Spanish, Chinese

**References**

1. Spanish/German: www.corpuspages.eu

Spanish/English: www.corpuspaens.eu

Spanish/Chinese: www.corpuspaches.eu

2. https://github.com/michaeljlang/PaCorEs-Splitter
3. LF-Aligner: https://sourceforge.net/projects/aligner/

Vecalign : https://github.com/thompsonb/vecalign/

Bertalign : https://github.com/bfsujason/bertalign/

# Concessive markers *although* and *aunque*: A case study based on the English-Spanish parallel corpus P-ACTRES 2.0

*Noelia Ramón (Universidad de León, Spain)*

Concession can be defined as a "relation that joins two clauses or units in a potential or apparent contradiction." (Taboada & Gómez-González, 2012, p. 18). Subordinate clauses of concession (or concessive clauses) express an idea that suggests the opposite of the main clause, thus involving a high degree of complexity at different levels: cognitive complexity, as it includes contrast between two propositions (Salkie & Oates, 1999), and syntactic complexity, as it is realized by subordinate clauses (Hasselgård, 2010, 2024). Concessive relations are not limited to a restricted number of markers but can be conveyed by a wide range of different discursive elements in different languages.

From a cross-linguistic perspective, concessive subordination enhances complexity, including shifts between coordination and subordination, and between phrase and clause subordination (Hasselgård, 2024). Previous studies of adverbial clauses of different types, including concessive clauses, suggest that there are differences in frequency, use and registers (Taboada & Gómez-González, 2012; Gast, 2019; Dupont, 2021). This paper will focus on the use of the most common concessive subordinators in English and Spanish - *although* and *aunque* - to investigate their frequency, use and register through their translations. The study will compute their mutual correspondence (MC) value (Altenberg, 2002), identifying thus the percentage of translations of *although* by *aunque* and vice versa. The aim is to show how information from a parallel corpus can provide insights into the various syntactic and positional options of the two most common concessive markers in English and Spanish via their translations in both directions. The empirical data will be extracted from the bidirectional parallel corpus P-ACTRES 2.0, in fictional as well as in non-fictional registers to reveal register differences intra-linguistically as well as inter-linguistically.

The English-Spanish parallel corpus P-ACTRES 2.0 is a large bidirectional translation corpus of contemporary texts, published in the year 2000 and later, including texts from a variety of different registers: fiction, non-fiction, newspapers, magazines and miscellanea. Today P-ACTRES includes nearly 6 million words of running text: 4.3 million words correspond to the subcorpus of English source texts and their corresponding Spanish translations, and 1.6

million words correspond to Spanish source texts and their English translations.

The English concessive marker *although* has been found to be more common in formal registers (Aarts, 1988; Biber et al. 1999, Huddleston & Pullum, 2002). A first approach to the analysis of *although* in P-ACTRES has yielded similar results, with a higher frequency in non-fictional texts (425.7 cases per million words) than in fictional texts (204.4 cases per million words). In contrast, in Spanish, *aunque* occurs more often in fiction (781.1 cases per million words) than in non-fiction (730.02 cases per million words), although the difference between the two registers is much less marked than in English. As for their degree of equivalence, preliminary results show a much higher number of translations of *although* as *aunque* than vice versa. This can be confirmed using the mutual correspondence (MC) value in both directions.

**Keywords:** contrastive analysis, concessive markers, English-Spanish, mutual correspondence.

## References

Aarts, B. (1988). Clauses of concession in written present-Day British English. *Journal of English Linguistics, 21*(1), 39-58. https://doi.org/10.1177/007542428802100104

Altenberg, B. (2002). Concessive Connectors in English and Swedish. In H. Hasselgård, S. Johansson, B. Behrens & C. Fabricius-Hansen (Eds.), *Information Structure in a Cross-Linguistic Perspective* (pp. 21-43). Rodopi. https://doi.org/10.1163/9789004334250_003

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

Dupont, M. (2021). Conjunctive markers of contrast in English and French. From syntax to lexis and discourse. John Benjamins.

Gast, V. (2019). A corpus-based comparative study of concessive connectives in English, German and Spanish. In: O. Loureda, I. Recio Fernández, L. Nadal & A. Cruz (Eds.), *Empirical studies of the construction of discourse* (pp.151-192). John Benjamins. https://doi.org/10.1075/pbns.305.06gas

Hasselgård, H. (2010). *Adjunct adverbials in English*. Cambridge University Press.

Hasselgård, H. (2024). Concessive subordination in English and Norwegian. *Languages in Contrast, 24*(1), 109-132. https://doi.org/10.1075/lic.00037.has

Huddleston, R. & Pullum, G.K. (2002). The Cambridge Grammar of the English Language. Cambridge University Press.

Salkie, R. & Oates, S.L. (1999). Contrast and concession in French and English. *Languages in Contrast*, *2*(1), 27-56.

Taboada, M. & Gómez-González, M.Á. (2012). Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences, 6*(1-3), 17-41. https://doi.org/10.1558/lhs.v6i1-3.17

# A Corpus-Driven Synchronic Analysis of the [ɛə] Diphthong Group: sounds and Spellings in the Nineteenth-Century Vernacular

*Nadia Hamade Almeida (Universidad Camilo José Cela, Spain)*

The Lancashire dialect has been widely represented in regional literature. Regional literature is commonly categorized into two types: dialect literature and literary dialect. Dialect literature refers to works that are entirely written in a non-standard linguistic variant. As a result, this category is mainly aimed at readers who are capable of reading and understanding the represented linguistic variant. Alternatively, literary dialect texts are predominantly written in standard English or a prestigious variety, except for the characters' dialogues, which depict a particular regional variant. One of the most salient characteristics of this form of literary representation is the use of non-standard spelling, derived from semi-phonetic transcriptions of standard English. For instance, the use of <ee> and <aw> to suggest the monophthongs [i:] and [ɔ:].

Literary-dialect texts are appropriate tools for researchers in dialect study (Ruano-García 2007: 111; Beal 2011: 204). A thorough examination of non-standard spellings provides phonological insight into a regional variety at a particular point in time. For this reason, the present study is built upon these texts to explore the Lancashire vernacular. Since providing a comprehensive view of the Lancashire dialect is beyond the scope of this article, this study focuses on the sounds and spellings related to the lexical group SQUARE, according to the classification Wells (1982:155) provides for terms associated with the standard diphthong [ɛə].

This article aims to outline and explain the different dialect pronunciations and the possible coexistence of sounds within the same lexical set, considering historical and sociolinguistic factors. For this purpose, a corpus comprising nineteen literary-dialect works, specifically selected from texts composed between 1850 and 1900, was assembled and analyzed manually. Since the dialect is principally represented in the characters' speech, this study centers on these dialogues, without overlooking the rest of the plot as it offers useful information about the context and the characters.

The various non-standard spellings associated with SQUARE were used as initial resource and then attributed to their corresponding sounds in the Lancashire dialect. In this regard, García-Bermejo Giner (1999: 252) argues that a comparison between standard and non-standard spelling is invaluable when

conducting phonological research through literary dialect texts. Once the deviant spellings related to [ɛə] and their corresponding sounds are gathered, this paper aims to elucidate the underlying reasons for those realizations, grounded in diachronic evolution of each one. This research undertakes both a quantitative and a qualitative analysis. The former aims to show the frequency index of the deviant spellings in the corpus, while the latter provides an insight into the sounds and spellings related to SQUARE.

The results of this research reveal the coexistence of two dialect sounds, [ɪə] and [ɜ:], within the SQUARE lexical group. The data indicates that the first pronunciation was once a common sound, widespread across Lancashire, but began to exhibit a regressive tendency during the nineteenth century. The second realization may signify a stereotypical form associated with the vernacular variant of Lancashire. Despite focusing on the dialect variations of a single RP sound, the present research contributes to broader dialect studies by using literary-dialect texts to trace phonological changes and uncover the historical and sociolinguistic factors that influence regional dialect evolution.

**Keywords**: Lancashire dialect, literary dialect, SQUARE lexical set, deviant spellings, dialect sounds.

**References**

Beal, J. (2011): English in modern times, London, Hodder Education

García-Bermejo Giner, F. (1999): "Methods for the linguistic analysis of early modern English literary dialects", in P. Alonso (Ed.) (1999) Teaching and research in English and linguistics, León, Celarayn: 249-266.

Ruano-García, J. (2007): "Thou'rt a strange fille: A possible source for 'y-tensing' in seventeenth-century Lancashire dialect?", Sederi, 17, 109-127.

Wells, J. C. (1982): Accents of English, Cambridge, Cambridge UP.

# Presentación del Corpus chino-español PaCheS

*Irene Doval (Universidad de Santiago de Compostela, Spain)*

El objetivo de esta presentación es dar a conocer PaCheS, un corpus paralelo en línea chino-español y sus funcionalidades actuales (www.corpuspaches.eu). Este corpus forma parte del proyecto PaCorES, una colección de corpus paralelos bidireccionales con el español como lengua central.

En primer lugar, se presentará la arquitectura general de la colección PaCorES, junto con los principios que la sustentan: calidad textual, multifuncionalidad, reusabilidad, accesibilidad, facilidad de uso y retroalimentación de los usuarios.

A continuación, se abordarán aspectos clave de la compilación del corpus PaCheS, incluyendo su composición y el preprocesamiento de los datos. Se prestará especial atención a la segmentación, al alineado automático de los textos, detallando la selección de la herramienta utilizada, estadísticas básicas del proceso y los principales fenómenos observados en la revisión de la alineación.

Por último, se describirán las opciones de búsqueda disponibles en la interfaz web del corpus, las opciones de visualización de resultados y las mejoras implementadas para optimizar y ampliar las funcionalidades del corpus. Finalmente, se esboza el desarrollo futuro del corpus.

**Palabras clave:** corpus paralelo, chino, español, lingüística de corpus

## Referencias

Doval, I., & Sánchez Nieto, M. T. (Eds.). (2019). Parallel corpora in focus: An account of current achievements and challenges. *In Parallel corpora for contrastive and translation studies* (pp. 1-18). De Gruyter. https://doi.org/10.1515/9783110643574-001

Liu, L., & Zhu, M. (2023). Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, *38*(2), 621-634.

Lefer, M. A. (2021). Parallel corpora. In *A practical handbook of corpus linguistics* (pp. 257-282). Cham: Springer International Publishing.

# Bridging Communication Gaps in EFL: AI-Driven Corpus Development and Analysis for Medical Discourse[1]

*Olga Freimane (University of Latvia)*

En los últimos años, los Grandes Modelos de Lenguaje (LLM) han tenido un enorme éxito en diversos campos de investigación. La arquitectura básica de la mayoría de los Grandes Modelos de Lenguaje (LLM) se basa en los denominados tranformers (Vaswani et al., 2017). Aunque los detalles exactos de la arquitectura de GPT-4 aún no están bien delineados, sabemos que el modelo subyacente es un transformers entrenado capaz de predecir el siguiente token dentro de una secuencia (OpenAI, 2024).

Esta investigación específica pretende explorar las posibilidades que ofrece la inteligencia artificial y, gracias al apoyo de los Grandes Modelos de Lenguaje (LLM) y del modelo GPT-4, se está intentando proceder a la anotación pragmática de diálogos en lengua española (corpus DiEspa - Diálogos en Español) (ParlarItaliano, 2008). En la base del proyecto se encuentra el esquema de anotación de diálogos Pr.A.T.I.D. (Pragmatic Annotation Tool for Italian Dialogues) (Savy, 2010), que en el momento de su creación requería el uso de un software de anotación manual (XGate), ahora obsoleto. Dado que ya no es posible anotar con el software XGate (Cutugno, D'Anna, 2006), se decidió intentar entrenar el modelo GPT-4 con el objetivo de acelerar la anotación pragmática de diálogos y refinar su memoria. Tradicionalmente, la anotación pragmática requiere una intensa actividad humana para identificar y clasificar los actos lingüísticos y los fenómenos pragmáticos complejos.

Mediante un enfoque iterativo, el modelo se adaptó para reconocer y anotar fenómenos pragmáticos específicos basados en el esquema de anotación de diálogos Pr.A.T.I.D. La metodología adoptada implica el preprocesamiento de los diálogos para adaptarlos a los requisitos del esquema Pr.A.T.I.D. (Castagneto, 2012), seguido del uso del modelo GPT-4 para la identificación preliminar de los moves dialógicos (Savy, Solís García, 2009) esperados que deben detectarse en el texto del diálogo. Posteriormente, las anotaciones generadas automáticamente fueron revisadas y corregidas por expertos humanos, garantizando así un alto nivel de precisión y fiabilidad. El

[1] The participation in the conference is funded within the Project "Internal and External Consolidation of the University of Latvia": Identification No. 5.2.1.1.i.0/2/24/I/CFLA/007.

trabajo de revisión permitió no sólo identificar posibles errores, sino corregirlos con el objetivo de perfeccionar el sistema y permitir que el modelo GPT-4 mejorara su rendimiento.

Además, se está probando la anotación mediante la actualización 4.5 de OpenAI. Se trata del nivel más avanzado de IA hasta la fecha, conocido internamente con el nombre en clave de

«Orion», que según la empresa puede proporcionar una interacción más natural capaz de entender las indicaciones del usuario de una forma «más humana». El nuevo modelo también incluye inteligencia emocional mejorada (sentiment analysis), lo que lo hace útil para tareas como mejorar la escritura, programar y resolver problemas prácticos, así como una reducción significativa de las «alucinaciones» en comparación con los modelos anteriores. De hecho, en versiones anteriores era frecuente encontrar «alucinaciones» en la aplicación del modelo Pr.A.T.I.D. con la introducción totalmente inventada de moves no previstas por la norma.

En conclusión, este estudio demuestra el potencial de los Grandes Modelos de Lenguaje (LLM) y del modelo GPT-4 para superar algunas de las limitaciones asociadas a la anotación pragmática manual. El objetivo es allanar el camino para futuras investigaciones sobre la aplicación de modelos generativos del lenguaje en otras áreas de la lingüística computacional y la posibilidad de ampliar este enfoque a otras lenguas y tipos de texto. Las implicaciones de este estudio se extienden más allá del campo de la anotación pragmática, sugiriendo nuevas direcciones para la integración de los Grandes Modelos de Lenguaje (LLM) en los procesos de análisis lingüístico.

**Keywords:** grandes modelos de lenguaje, anotación pragmática, GPT-4, diálogos

## References

Castagneto Marina, (2012), “Il sistema di annotazione Pra.Ti.D tra gli altri sistemi di annotazione pragmatica. Le ragioni di un nuovo schema”, in ANNALI del Dipartimento di Studi Letterari, Linguistici e Comparati Sezione Linguistica, Università degli Studi di Napoli “L’Orientale”, Napoli, Italia.

Cutugno Francesco, D'Anna Leandro, (2006) “XGate e XRG: strumenti per l'editing visuale, l'interrogazione e il benchmarking di annotazioni

linguistiche XML", Dipartimento di Fisica - Gruppo NLP, Università 'Federico II' di Napoli, Italia, Dipartimento di Linguistica e Letteratura, Università di Salerno, Italia.

OpenAI, (2024), Arxiv, URL: https://arxiv.org/abs/2303.08774

ParlarItaliano, Studium Dipsum, (2006), URL: https://parlaritaliano.studiumdipsum.it/it/792-corpus-diespa-dialogos-en-espanol

Savy Renata, (2010) "Pr.A.T.I.D.: a coding scheme for pragmatic annotation of dialogues", Dipartimento di Linguistica e Studi Letterari dell'Università di Salerno, Fisciano (Salerno), Italia.

Savy Renata, Solís García Inmaculada (2008), "Strategie pragmatiche in italiano e spagnolo a confronto: una prima analisi su corpus", Università degli Studi di Salerno, Fisciano (Salerno), Italia.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N.

Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need." In: Advances in Neural Information Processing Systems. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://arxiv.org/abs/1706.03762

# Merck DE-ES: un corpus para el estudio de la comunicación médica en situaciones comunicativas simétricas y asimétricas

*María Teresa Sánchez Nieto (Universidad de Valladolid, Spain)*

El Parallel Corpus of German and Spanish, en adelante PaGeS (Doval Reixa et al., 2019) , es un corpus paralelo bilingüe bildireccional que, junto con PaEnS, PaCheS y PaFreS integra la colección PaCorES (Parallel Corpus of Spanish), todos ellos accesibles en línea sin restricciones a través de www.pacores.eu. Cada uno de los corpus bilingües está conformado por un corpus nuclear y una serie de "suplementos". Mientras que los corpus nucleares son amplias colecciones de textos de prosa contemporánea (ficción y no ficción), los suplementos son corpus bilingües paralelizados de menor tamaño que los corpus nucleares y que representan usos especializados de la lengua (Doval & Sánchez Nieto, 2026). Actualmente PaGeS cuenta con tres suplementos: TED Talks DE-ES (Cettolo et al., 2012), que representa el uso divulgativo del lenguaje científico; Europarl DE-ES (Graën et al., 2014), que representa el uso administrativo de la lengua; y OpenSubtitles DE-ES (Lison & Tiedemann, 2016), que testimonia el uso de la lengua en los subtítulos.

Este trabajo presenta el proceso de compilación de un nuevo suplemento para PaGeS: el suplemento Merck DE-ES. Se trata de un corpus paralelo bilingüe a partir de las versiones alemana y española de los manuales Merck disponibles en línea (*Manual Merck versión para profesionales*, s. f.; *Manual MSD versión para público general*, s. f.; *MSD Manual Ausgabe für Patienten*, s. f.; *MSD Manual Profi-Ausgabe*, s. f.). Los manuales Merck son obras de referencia médica que se publicaron por primera vez en 1899 como una guía de bolsillo para médicos y farmacéuticos. Con el tiempo, se transformaron en un recurso en línea con versiones tanto para facultativos y otros profesionales de la salud (versión para profesionales) como para consumidores (versión para el hogar) (Bullers, 2017, p. 369; Tomes, 2021, p. 1). Los manuales ofrecen información médica detallada escrita en un lenguaje adaptado a cada uno de los públicos objetivo mencionados, con el fin de proporcionar información médica imparcial y accesible. Por lo tanto, el suplemento Merck DE-ES representará el uso especializado de la lengua en el ámbito médico en situaciones de comunicación simétricas (experto-experto, p. ej. en la comunicación entre médicos) y asimétricas (experto-lego, p. ej. en la comunicación médico-paciente).

Tras presentar brevemente la historia, características e interés de los manuales Merck para el proyecto PaGeS/PaCorES, explicitaremos los criterios de selección del material y de los metadatos necesarios para integrar el suplemento en la arquitectura de PaCorES. Asimismo, describiremos las características del *script* Merk2Text_info (Oliver, 2025), que permite la extracción y conversión automatizada de los datos a partir de las páginas web que contienen los textos alemanes y españoles que nos interesan, y, finalmente, los pasos del posprocesamiento de los datos para su indexación en PaGeS. Tras exponer las estadísticas principales del corpus, reflexionaremos sobre el carácter piloto de esta compilación, con el fin de valorar la posibilidad de compilar con esta técnica suplementos Merk para PaEnS (inglés-español), PaFrEs (francés-español) y PaChEs (chino-español).

**Referencias**

Bullers, K. (2017). Merck Manuals. Journal of the Medical Library Association, 104(4). https://doi.org/10.5195/jmla.2016.164

Cettolo, M., Girardi, C., & Federico, M. (2012). WIT3: Web inventory of transcribed and translated talks. *Proceedings of the 16th Annual Conference of the European Association for Machine Translation, EAMT 2012*.

Doval, I., & Sánchez Nieto, M. T. (2026). Parallel Corpora Spanish (PaCorES): A collection of multifunctional parallel corpora. *RESLA. Revista Española de Lingüística Aplicada / Spanish Journal of Applied Linguistics.*

Doval Reixa, I., Fernández Lanza, S., Jiménez Juliá, T. E., Liste Lamas, E., Lübke, B., Doval Reixa, I. (ed. lit. ), & Sánchez Nieto, M. T. (ed. lit. ). (2019). Corpus PaGeS: A multifunctional resource for language learning, translation and cross-linguistic research. En *Parallel Corpora for Contrastive and Traslation Studies: New resources and applications* (pp. 103-121). Amsterdem:John Benjamins,2019. https://dialnet.unirioja.es/servlet/extart?codigo=7225008

Graën, J., Batinic, D., & Volk, M. (2014). Cleaning the Europarl Corpus for Linguistic Applications. *Konvens 2014, Hildesheim, 8 October 2014 - 10 October 2014*, 1-7. https://www.zora.uzh.ch/id/eprint/99005/

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

*Manual Merck versión para profesionales*. (s. f.). Recuperado 16 de febrero de 2025, de https://www.merckmanuals.com/es-us/professional

*Manual MSD versión para público general*. (s. f.). Recuperado 16 de febrero de 2025, de https://www.msdmanuals.com/es/hogar

*MSD Manual Ausgabe für Patienten*. (s. f.). Recuperado 16 de febrero de 2025, de https://www.msdmanuals.com/de/heim

*MSD Manual Profi-Ausgabe*. (s. f.). Recuperado 16 de febrero de 2025, de https://www.msdmanuals.com/de/profi

Oliver, A. (2025). *Merk2Text_info*.

Tomes, N. (2021). Not just for doctors anymore": How the merck manual became a consumer health "bible. *Bulletin of the History of Medicine*, *95*(1). https://doi.org/10.1353/bhm.2021.0000

# Race with the machines: Using corpora to assist LLM-based machine translation?

*Maria Teresa Musacchio (University of Trieste, Italy)*

*Giuseppe Palumbo (University of Trieste, Italy)*

Despite translators' concerns for their profession in a future world where human translation will mainly consist of revision of machine output, the underlying problem remains: ensuring translation quality, no matter whether it is the outcome of human or automated processes. Automated translation tasks can today be performed both by neural machine translation systems and by AI-based LLMs. These can be trained with in-context learning (ICL) strategies, which have been shown to improve the translation capabilities of LLMs (Lyu et al., 2023; Zhu et al. 2024).

This study aims to evaluate the translation performance of large language models (LLMs) when translating from English into Italian a Wikipedia entry on "consumer behaviour" under two general conditions: with and without the support of a comparable corpus in the target language. The comparable corpus comprises freely available, academically-oriented texts on consumer behaviour. The research investigates how access to relevant texts impacts an LLM's ability to accurately render terminology and appropriate phraseology in the target language.

The methodology involves four distinct translation processes. In the first (Process 1), an LLM generates a translation without any contextual exposure to our comparable corpus. In Process 2, the translation is carried out by a freely available neural machine translation service. In Process 3, the text is translated with Wikipedia's (2025) own machine translation system. In Process 4, the same LLM as in Process 1 is used, but this time the LLM is asked to translate the text with the assistance of our comparable corpus in the target language. More specifically, the comparable texts are fed into the system as one large corpus, and – following a few-shot strategy (Garcia et al. 2023) – the system is prompted to treat it as a set of terminological and phraseological references to draw on. Expert evaluators then assess the four translated texts on the basis of the reduced scorecard available in ISO 5060:2024 *Translation services—Evaluation of translation output—General guidance*. In particular, evaluation of terminology is carried out with reference to published specialized parallel texts in the domain of marketing.

The results are analyzed to determine whether the use of comparable corpus enhances the LLM's translation accuracy, particularly regarding terminology and nuanced phrasing. By comparing the four target texts, the study highlights the importance of contextual input in improving the quality of AI-generated translation.

**Keywords:** neural machine translation, AI-generated translation, LLMs, comparable corpus, English, Italian

**References**

Garcia, X., Bansal, Y., Cherry, C., Foster, G., Krikun, M., Johnson, M., & First, O. (2023). The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, https://arxiv.org/abs/2302.01398.

Lyu, C., Xu, J., & Wang, L. (2023). New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*, https://longyuewang.com/pdf/New_Trends_in_Machine_Translation_using_LLMs.pdf.

Wikipedia, (2025). Aiuto:Strumento Traduzione voci, https://it.wikipedia.org/wiki/Aiuto:Strumento_Traduzione_voci (last accessed on 14 February 2024)

Zhu, S., Cui, M., & Xiong, D. (2024, May). Towards robust in-context learning for machine translation with large language models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 16619-16629), https://aclanthology.org/2024.lrec-main.1444.pdf.

# 3. Consideraciones interlingüísticas e interculturales en textos generados por IA / Cross-linguistic and cross-cultural considerations in AI-generated texts

# Inteligencia Artificial y Fake News: Análisis Lingüístico, Sesgo de Género y Ética en la Comunicación Digital de los Medios

*Julia Mary Scilabra (Universidad Complutense de Madrid, Spain)*

*Dario Russo (Universidad de Málaga, Spain)*

The digital landscape is a privileged space for the manipulation and subsequent dissemination of false or mystified content, often amplifying gender bias through the misuse of terminology. Recent studies have shown that content mystification and use of gender bias are perpetrated, if not reinforced, when artificial intelligence tools are used to generate, moderate, and disseminate content, thereby contributing to the construction of narratives that perpetuate stereotypes and misinformation (Minucci et al., 2022; Al-Asadi & Tasdemir, 2022; Nazar & Bustam 2020; Marchetti, 2019; Biju &. Gayathri, 2023; etc.). These dynamics highlight the need for a critical approach in the use of such technologies and for greater ethical and methodological attention. Specifically, particular attention should be paid to training data selection and processing and to terminological choices in order to avoid the reinforcement of bias and discrimination.

By combining a linguistic and cross-linguistic approach with technical-theoretical aspects of digital communication practices and AI ethics, the present exploratory study will adopt a mixed method aimed at providing deeper understanding of the phenomenon. Specifically, a qualitative approach will be used to analyze the linguistic and discursive dynamics underlying the use of gender-biased terms in AI-generated content; while the quantitative approach will provide measurable data on the frequency and distribution of these phenomena.

A comparative analysis will be conducted on the selected online articles focusing on topics identified by Google Trends as the most searched in 2024 in the UK and the US; the mechanisms through which the choice of terms can influence public perception, reinforce gender biases or, conversely, promote more inclusive and ethically conscious communication will be examined.

Indeed, the two aforementioned countries, the United States and the United Kingdom, share the same most frequent search keys in the year 2024: the newly elected U.S. president, Donald Trump, and Imane Khelif, an Algerian boxer who took part in the 2024 Paris Olympics. Suffering from hyperandrogenism, over the course of the Olympics, Khelif's figure has been

the center of much discussion and speculation, especially following rumors regarding her possible identification as a transgender athlete. The interest raised by these two issues is a significant element for the current analysis, as it highlights the convergence of social-political dynamics and issues related to gender representation, reflecting major research trends and social debates in the U.S. and British contexts.

The newspapers considered in the study were identified through the Digital News Report 2024 from the Reuters Institute for the Study of Journalism | University of Oxford. The study sample was restricted to The New York Times, The Washington Post, The Guardian, and Daily Mirror. In fact, the use of AI by these newspapers was verified through tools of advanced source code analysis - including AI-generated code - which also detect instances of plagiarism in case of code alteration. The articles examined via Google Search's advanced filters were limited to the period between July 26 and August 11, 2024, i.e., the days of the Olympic event. The most relevant content in that time span for the case study was identified through the commands *allintitle:Imane Khelif site:nameofthewebjournal*.

The ultimate goal of the present study is therefore to promote ethical and terminological reflection on how words and algorithms are intertwined in shaping social perceptions of gender and how disinformation and misinformation might be exploited or purposely fabricated for propaganda. The study thus aims to spark a more informed and informed debate on these complex dynamics.

**References**

Nakov, P. and Da San Martino, G. (2021). *Fake News, Disinformation, Propaganda, and Media Bias*. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21). Association for Computing Machinery, New York, NY, USA, 4862–4865. https://doi.org/10.1145/3459637.3482026

O'Connor, S., & Liu, H. (2023). Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & Society*, *39*(4), 2045–2057. https://doi.org/10.1007/s00146-023-01675-4

Marinucci, L., Mazzuca, C., & Gangemi, A. (2022). Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. *AI & Society*, *38*(2), 747–761. https://doi.org/10.1007/s00146-022-01474-3

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). *Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies*. AIS Electronic Library (AISeL). https://aisel.aisnet.org/acis2020/27/?utm_source=aisel.aisnet.org%2Facis2020%2F27&utm_medium=PDF&utm_campaign=PDFCoverPages

Al-Asadi M., Trasdemir R., (2022). Using Artificial Intelligence Against the Phenomenon of Fake News: A Systematic Literature Review in Studies in Computational Intelligence. *Combating Fake News with Computational Intelligence Techniques, 1001, 39-54*

Biju, P. R., & Gayathri, O. (2023). Self-Breeding Fake News: Bots and Artificial Intelligence Perpetuate Social Polarization in India's Conflict Zones. The International Journal of Information, Diversity, & Inclusion, 7(1/2), 1–25. https://www.jstor.org/stable/48731169

Marchetti G., (2020) Le fake news e il ruolo degli algoritmi. Media Laws. https://www.medialaws.eu/wp-content/uploads/2020/03/1-2020-Marchetti.pdf

Zimmer, F. Scheibe, K., Stock, M., & Stock, W. G. (2019). Fake News in Social Media: Bad Algorithms or Biased Users?. JOURNAL OF INFORMATION SCIENCE THEORY AND PRACTICE, 7(2), 40-53, https://doi.org/10.1633/JISTaP.2019.7.2.4

# Dinos cómo traduces *Chair* y te diremos quién eres. Análisis del sesgo de género en traducciones poseditadas

*Cristina Toledo-Báez (IUITLM, Universidad de Málaga, Spain)*

*María Jesús García Serrano (IUITLM, Universidad de Málaga, Spain)*

Las tecnologías del lenguaje y la Inteligencia Artificial desempeñan un papel fundamental en nuestro día a día. Dentro del campo de las tecnologías del lenguaje y la Inteligencia Artificial aplicada al ámbito de las Humanidades, la traducción automática acarrea limitaciones en cuanto al empleo de un lenguaje inclusivo, lo que apoyan, entre otros, Savoldi et al. (2024). Diversos autores, entre ellos Savoldi (2021) y Attanasio (2023), han demostrado que los sistemas de traducción automática se decantan por las formas masculinas. Partiendo de este punto de partida, el objetivo del presente trabajo es evaluar las connotaciones de género en la traducción automática y en la posedición humana de tres textos traducidos automáticamente.

Nuestro estudio se enmarca en el proyecto NEUROTRAD (B1-2020_07) de la Universidad de Málaga (Toledo Báez, 2024a; Toledo Báez, 2024b; Toledo-Báez y García Serrano, 2025/en prensa), en el seno del cual se llevó a cabo entre 2022 y 2023 un estudio empírico con 29 traductores/poseditores profesionales que poseditaron tres tipos de traducciones distintas (automática, humana y automática poseditada) que posteriormente se evaluaron con una taxonomía adaptada de MQM (Multidimensional Quality Metrics). En esta comunicación nos centraremos en la posedición del género en cada uno de los participantes centrándonos en los siguientes puntos:

1. Extracción en cada uno de los textos de todas aquellas palabras en inglés que se refieren a personas y que, al ser traducidas al español, puedan tener una connotación de género.
2. Evaluación del sesgo en las propuestas realizadas por DeepL, el traductor automático escogido para el estudio empírico de NEUROTRAD.
3. Comparativa de los datos demográficos de los participantes y su formación con la posedición del género.

Nuestro estudio aborda la tendencia que pueda establecerse entre datos demográficos concretos de los participantes con la no posedición o posedición incorrecta de sesgos de género, además de la necesidad de seguir investigando en la línea de la formación en posedición y la corrección del

género en textos traducidos automáticamente, para evitar así la perpetuación de los sesgos de género y su refuerzo en la sociedad.

**Palabras clave:** Posedición, sesgo de género, evaluación del error, lenguaje inclusivo

**Referencias**

Attanasio, G., Plaza-Del-Arco, F. M., Nozza, D., & Lauscher, A. (2023, 18 octubre). A Tale of Pronouns: Interpretability Informs Gender Bias Mitigation for Fairer Instruction-Tuned Machine Translation. arXiv.org. https://arxiv.org/abs/2310.12127

GITT 2024 - DeBiasByUS. (s. f.). https://sites.google.com/tilburguniversity.edu/gitt2024/more/debiasbyus

Piergentili, A., Savoldi, B., Negri, M., & Bentivogli, L. (2024, 14 mayo). Enhancing Gender-Inclusive Machine Translation with Neomorphemes and Large Language Models. arXiv.org. https://arxiv.org/abs/2405.08477

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. Transactions Of The Association For Computational Linguistics, 9, 845-874. https://doi.org/10.1162/tacl_a_00401

Toledo-Báez, C. & García-Serrano, María J. (en prensa). ¿Es errar solo de humanos? Análisis de errores en traducción automática neuronal y posedición. Comares.

Toledo-Báez, C. (2024A). Paridad humano-máquina en traducción automática neuronal y posedición: Un estudio empírico desde la traducción profesional. Lebende Sprachen 2024; 69(2): 1–29. DOI: 10.1515/ les - 2024 – 0003

Toledo-Báez, C. (2024B). Propuesta de enseñanza-aprendizaje de posedición con la paridad humano-máquina. Tecnologías lingüísticas multilingües: desarrollos actuales y transición digital, pp 79-98, Comares: Granada

# Velo en flor', 'flower veil' o 'the veil of flor': ¿cómo traduce los culturemas la IA?

*Juan Pedro Morales Jiménez (GIRTraduvino, Universidad de Valladolid, Spain)*

La instauración de la Inteligencia Artificial en las empresas vitivinícolas es cada vez mayor dada la variedad de recursos y utilidades que pueden desempeñar. Consecuentemente, el proceso de traducción también se ha visto influenciado al surgir nuevas fases como la posedición. El problema radica en que las traducciones generadas por la IA no siempre se ajustan a la corrección o a la adecuación al contexto sin la revisión por un humano.

Por ello, en el presente análisis partimos de esa premisa y nos planteamos comprobar si las diversas herramientas traducían adecuadamente el discurso vitivinícola y, en particular, el del vinagre D. O .P. La lengua de la vid y el vino se caracteriza por una nutrida terminología relacionada con las metáforas y la cultura. Sin un conocimiento del contexto o del sector, la traducción de algunas unidades léxicas puede derivar en un error. Los culturemas son una de las principales complejidades de la Traductología y requieren un cuidado trasvase según el contexto. Un traductor humano tiene la capacidad de adaptarse a las necesidades comunicativas, pero la IA necesita un pautado muy concreto para lograrlo.

En la presente investigación tomamos la hipótesis de que las nuevas tecnologías aplicadas a la traducción cultural, gastronómica y vitivinícola no son capaces de conseguir la adecuación de los culturemas en diversos contextos. Gracias al corpus paralelo del macrocorpus sobre vinagre *Aceticorpus*, analizamos cómo se trataban los culturemas presentes. Seleccionamos aquellas fichas técnicas de vinagres con D. O. P. traducidas para generar textos metas realizados por traductores automáticos (Google Translate y DeepL) y por IA (Chat GPT-4, Gemini Flash 2.0 y Copilot GPT-4). Aunque los dos primeros no son recursos conversacionales, son unos de los principales traductores *online* y resulta de gran interés conocer si la calidad se asemeja. En el resto de herramientas, es necesario establecer una conversación y, de dicho modo, se puede imponer unas instrucciones. Les propusimos mediante *prompts* secuenciales que: 1. tradujesen una serie de textos al inglés: «¿puedes traducir esto al inglés?»; 2. que lo adaptasen a una persona conocedora de la cultura: «¿consideras que es adecuado para una persona que conoce la cultura de _____»? En caso contrario, modifica la traducción» y,

finalmente, 3. que redactase el texto para una persona ajena a ella: «¿Podrías adaptarlo para una persona que no conoce la cultura?».

Contamos con un total de 3 versiones de cada texto generadas por cada recurso, a lo que se suma el texto meta originado en las bodegas. Tras extraer los culturemas más característicos mediante Sketch Engine, aplicamos una metodología analítico-contrastiva para conocer qué técnicas son más empleadas según el contexto y la herramienta. Preliminarmente, no hemos encontrado diferencias sustanciales dado un predomino de técnicas extranjerizadoras en los tres contextos contemplados. Concluimos, *a priori*, que las herramientas consultadas no son capaces de modificar su discurso adecuadamente para generar una comprensión de los culturemas. Por ende, la IA no mantiene unos estándares de adecuación ni de calidad que permita su aplicación en contextos culturales específicos sin revisión humana. Para futuros estudios sería interesante cambiar la región geográfica del corpus para comprobar si la metodología efectuada tiene resultados similares.

**Referencias**

Biber, D., Conrad, S., y Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

Caballero, R., Suárez-Toste, E., & Paradis, C. (2019). *Representing Wine. Sensory Perceptions, Communication and Cultures*. John Benjamins Publishing Company. https://doi.org/10.1075/celcr.21.

Kenning, M. M. (2010). What are parallel and comparable corpora and how can we use them? En A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 487-500). Routledge, Taylor & Francis Group.

Luque Janodet, F. (2022). Hacia una caracterización del discurso de la cata de vino como lengua de especialidad. *Onomázein Revista de lingüística filología y traducción*, *58*, 1-17. https://doi.org/10.7764/onomazein.58.01.

Paciaroni, E. C. (2024). La comunicación del Patrimonio Gastronómico Italiano y su traducción por la Inteligencia Artificial: el Caso del Vinagre Balsámico Tradicional de Módena D. O. P. [Trabajo de Fin de Máster, Università Degli Studi di Padova]. Thesis and dissertation Padua Archive. https://hdl.handle.net/20.500.12608/70289.

Pascual Cabrerizo, M. y Martínez Lanzán, G. (2024). La eficacia de la inteligencia artificial en la traducción de fichas técnicas de vinos. En

Almudena Barrientos Báez, Alba María Martínez-Sala y David Caldevilla Domínguez (Eds), *Caer en la red con Inteligencia (Artificial)* (pp. 291-300). Marcial Pons.

# Cross-linguistic and cross-cultural considerations in AI-generated texts: A Case Study of EFL Iraqi Students

*Kamal Khazal Mohammd (Universidad de Valladolid, Spain; Imam Al Kadhum College, Iraq)*

The research paper discusses the cross-linguistic and cross-cultural factors in AI generated texts, touching their implications in effective communication, cultural appropriateness, and linguistic accuracy. Given the increasing dependence on AI tools for content creation in education, work, and creativity, it is extremely relevant that one considers the question of how such tools find their way through and accommodate the complexities and diversities of languages and cultures. The present study investigates the cultural norms, idioms, and syntax structures by which AI-generated texts travel across languages, with a particular eye towards Arabic and English.

It adopts mixed-methods approaches to examine an AI-generated educational case study designed for Iraqi Arabic-speaking students learning English as a Foreign Language (EFL). The sample comprises 50 educational texts produced by GPT-based AI systems. They are analyzed for linguistic accuracy and cultural relevance and in the extent to which they relate to expectations from the collective audience with respect to language and culture. Texts are evaluated by a panel of bilingual linguists and cultural experts, who identify patterns of cultural misalignment, inappropriate tone, and potential grammatical errors that could create misunderstanding or cultural misappropriation.

Preliminary investigations also suggest that while AI systems are achieving remarkable feats in generating grammatically correct sentences, they fall short when it comes to producing translations and conveying nuanced cross-cultural meanings in culturally appropriate ways. They literally translate some idiom phrases, offer insufficient localization of examples, and show no sensitivity to cultural taboos or norms. These limitations affirm the need of an incorporation of cross-linguistic and cross-cultural frameworks into AI training models for inclusion and effective communication. Recommendations to extend AI text generation systems merge with the study conclusion and reinforce cultural corpora integration, collaboration with linguistic researchers, and development of algorithms with adaptive features prioritizing context-sensitive language production. Thus the contribution to the broader discussion on ethical and practical implications of AI in multilingual and multicultural contexts is made by

these findings toward providing actionable points to developers, educators, and policymakers.

# 4. El papel de los modelos de lenguaje de gran escala (LLMs) en la lingüística de corpus / The role of large language models (LLMs) in corpus linguistics

# Capitalizing on genre-based corpora with the use of AI-powered research tool Notebook LM

*Belén Labrador (Universidad de León, Spain)*

Syntagmatic relations have always been at the core of corpus linguistics, where the most important element to study is co-text - mainly collocates and patterns identified in concordance lines. Without abandoning this approach, the advent of AI has opened new possibilities to exploit corpora in different ways beyond the features of the corpus browsers or concordancers. AI-powered research tools have broadened perspectives by providing a deeper understanding of corpora that includes a more holistic scope. This presentation aims to explore the potential of a generative AI (GenAI) application, Notebook LM, in providing information that complements the data retrieved by traditional corpus-analysis toolkits. Notebook LM has been mainly used so far in education to generate podcasts with a feature called audio overview (Alonso-Guisande, & López-Fraile, 2024, Mehta et al., 2024). In the present paper, some other features of Notebook LM have been tried to support the initial hypothesis claiming that by combining corpus and GenAI approaches, we can gain a more comprehensive understanding of specific genre-based texts and therefore of the language used. While it has been claimed that "one of the main advantages of corpora is that we know exactly the domain of texts from which the corpus data is derived, something that we cannot track from current large language models underlying applications like ChatGPT" (Crosthwaite & Baisa, 2023), the use of GenAI applications like Notebook LM overcomes this shortcoming, as it only makes use of the texts that are uploaded as the sources of data, which can be our corpora.

A comparable genre-based English-Spanish corpus has been used as a case study to illustrate a proposed research methodology based on this AI-based tool. This corpus contains online cheese descriptions (400 texts in Spanish and 600 in English amounting to a similar number of words: 121,461 words in Spanish and 111,871 in English). It has been fed as a source of texts to Notebook LM and the different functions available have been used to summarize and explain the contents of the corpus and retrieve key themes, such as historical significance, Protected Designation of Origin, regionality and terroir, production methods, variety of milk types, importance of maturation, use of vegetarian rennet, flavour profiles and unique characteristics. As well as the features offered in the Notebook LM interface, the power of prompting is

also shown with some specific queries, whose answers seem to indicate that GenAI tools are more adequate for qualitative than quantitative analysis. The writing conventions in each language in this particular domain are compared and the main differences found in terms of style and language use point towards more concise and informative descriptions in English whereas the Spanish texts make use of more subjective and evocative language, often including more cultural context, even anecdotes, and a more enthusiastic tone. These findings can assist in second-language writing and translation training or practice. The presentation concludes by acknowledging the fact that Notebook LM provides complementary affordances which, combined with other corpus-analysis applications, constitute a powerful tool to capitalize on genre-based corpora.

**Keywords:** Notebook LM, Generative AI applications, genre-based corpus, comparable corpus, prompts.

**References**

Crosthwaite, P. & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! Applied Corpus Linguistics 3(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066

Mehta, N., Agrawal, A., Benjamin, J., Mehta, S., MacNeill, H., & Masters, K. (2024). Pedagogy and generative artificial intelligence: Applying the PICRAT model to Google NotebookLM. Medical Teacher, 1–3. https://doi.org/10.1080/0142159X.2024.2418937

Alonso-Guisande, M.A., & López Fraile, L.A. (2024). Herramientas de inteligencia artificial generativa aplicadas en la producción de podcasts. Edu Review, 12(2), 19-32. https://doi.org/10.62701/revedu.v12.5409

# El corpus ROBOT-TALK para el reconocimiento del origen robótico de textos en español

*Ana M.ª Fernández-Pampillón Cesteros (Universidad Complutense de Madrid, Spain)*

*Lara Alonso Simón (Universidad Complutense de Madrid, Spain)*

ROBOT-TALK es un corpus monitor comparable de textos humanos en español y su contrapartida escrita por grandes modelos generativos del lenguaje (en adelante «modelos del lenguaje»). Su objetivo es permitir el estudio de posibles rasgos lingüísticos diferenciadores entre textos generados automáticamente y los generados por las personas en el marco del PID2022-140897OB-I00 (https://www.ucm.es/robottalk/).

Los modelos del lenguaje producen textos gramaticalmente correctos y con alto nivel de coherencia, por lo que es difícil distinguirlos de los humanos (Uchendu, Le y Lee, 2023). Esto plantea desafíos de alto impacto político, económico y social cuando se usan con fines maliciosos (Pizarro, 2019; Pavlyshenko, 2022; Cardenuto, Yang, Padilha, Wan, Moreira, Li, Wang, Andaló, Marcel y Rocha, 2023; Crothers, Japkowicz y Viktor, 2023). Entonces, resulta clave conocer si el autor es una persona o una máquina (Maloyan, Nutfullin y Ilyushin, 2022). Así, ROBOT-TALK proporciona el primer recurso lingüístico en español para el reconocimiento de autoría humana vs. «robótica» de textos.

Contiene textos humanos y contrapartidas «robóticas». Cubre tres géneros de diferentes niveles de formalidad en la lengua escrita: artículos científicos, noticias y reseñas. Cada par de textos, de longitud similar, trata el mismo tema. Los modelos del lenguaje se seleccionaron aplicando dos criterios: que generen textos de muy alta calidad y sean accesibles mediante API o una interfaz. Se recogen muestras de Claude de Anthropic, Falcon de Technology Innovation Institute, ChatGPT-3.5-turbo y ChatGPT-4 de OpenAI, Gemini de Google y Mixtral-8x7B-Instruct-v0.1 de Mixtral AI. El periodo de recogida abarca desde julio de 2023 hasta febrero de 2025. La tabla 1 muestra la distribución por tipo de autor y género textual.

| corpus | huma | bard | clau | g35t | gpt4 | mxit | total por |
|---|---|---|---|---|---|---|---|
| **artículos** | 144 | 90 | 0 | 90 | 90 | 90 | 504 |
| **noticias** | 171 | 171 | 60 | 111 | 151 | 111 | 775 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **reseñas** | 160 | 160 | 65 | 95 | 160 | 95 | 735 |
| **total por autor** | 475 | 421 | 125 | 296 | 401 | 296 | **2014** |

Tabla 1. Composición del corpus

Actualmente, el corpus consta de 2014 textos: 475 humanos (144 artículos, 171 noticias y 160 reseñas) y 1539 generados por modelos del lenguaje.

El método de construcción consta de cuatro pasos:

(1) búsqueda del texto humano, asegurando que la autoría es humana;

(2) almacenamiento del texto humano en formato xml para describir los metadatos y la estructura del contenido;

(3) generación del texto robótico comparable mediante *prompts* que determinan que el contenido tenga el mismo registro, longitud y temática que el texto humano comparable; y

(4) almacenamiento del texto robótico en formato xml con los metadatos y la estructura del contenido.

El etiquetado en xml de los textos del corpus permite su consulta con cualquier herramienta de análisis textual (ej. SketchEngine) que soporte este estándar de marcado. En este sentido, ROBOT-TALK se ha utilizado con la herramienta SketchEngine para realizar (1) un análisis lingüístico profundo para encontrar los rasgos más salientes que caracterizan los textos generados por los modelos de lenguaje; (2) un análisis estadístico de rasgos lingüísticos propios de los modelos del lenguaje frente a un posible estilo general humano en español (Alonso Simón, Fernández-Pampillón Cesteros, Fernández Trinidad y Márquez Cruz, 2024); y (3) la construcción de clasificadores automáticos binarios y multiclase basados en aprendizaje automático para distinguir textos róticos y humanos. El corpus no está todavía publicado, pero se puede consultar una muestra en https://www.ucm.es/robottalk/corpus-robot-talk.

**Palabras clave:** Corpus monitor, corpus comparable, corpus de texto en español, grandes modelos del lenguaje, identificación de textos automáticos.

**Referencias**

Alonso Simón, L., Fernández-Pampillón Cesteros, A.M., Fernández Trinidad, M. y Márquez Cruz, M. (2024). ¿Tienen GPT-3.5 y GPT-4 un estilo de escritura

diferente del estilo humano? Un estudio exploratorio para el español. *RAEL: Revista Electrónica de Lingüística Aplicada*, *23*, 34-54. https://doi.org/10.58859/rael.v23i1.666

Cardenuto, J. P., Yang, J., Padilha, R., Wan, R., Moreira, D., Li, H., Wang, S., Andaló, F., Marcel, S. y Rocha, A. (2023). The Age of Synthetic Realities: Challenges and Opportunities. *APSIPA Transactions on Signal and Information Processing*, *12*(1), 1–62. https://doi.org/10.1561/116.00000138

Crothers, E. N., Japkowicz, N. y Viktor, H. L. (2023). Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *arXiv:2210.07321*, Oct. 2023. https://doi.org/10.1109/ACCESS.2023.3294090

Maloyan, N., Nutfullin, B. y Ilyushin, E. (2022). DIALOG-22 RuATD Generated Text Detection. En *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"* (2022) (pp. 396–401). Moscú: RSUH. *arXiv.2206.08029*. https://doi.org/10.48550/arXiv.2206.08029

Pavlyshenko, B. M. (2022). Methods of Informational Trends Analytics and Fake News Detection on Twitter. *arXiv:2204.04891v1*. https://doi.org/10.48550/arXiv.2204.04891

Pizarro, J. (2019). Using n-grams to detect bots on Twitter, notebook for PAN at CLEF 2019. En L. Cappellato, N. Ferro, D. E. Losada, and H. Müller (Eds.), *CLEF 2019 Labs and Workshops, Notebook Papers*. https://ceur-ws.org/Vol-2380/paper_183.pdf

Uchendu, A., Le, T. y Lee, D. (2023). Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter, 25*(1), 1–18. https://doi.org/10.1145/3606274.3606276

# 5. Mejora de la enseñanza de idiomas con IA y corpus / Enhancing language teaching with AI and corpora

# Enseñanza de caracteres chinos en el contexto de la IA

*Wanruo Luo (Universidad de Salamanca, Spain)*

Este estudio aborda la contradicción entre la estandarización y la personalización en la enseñanza de caracteres chinos, construyendo un corpus multimodal de escritura (HCCD-2.0) y desarrollando un modelo de aprendizaje profundo para evaluar cuantitativamente el impacto de la inteligencia artificial en la adquisición de caracteres chinos por parte de hablantes no nativos. Integramos la base de datos HCDB de la Universidad de Pekín con una colección propia de 5000 muestras manuscritas, utilizando una versión mejorada de YOLOv5 para realizar detección de errores a nivel del trazo -orden del trazo/estructura-, combinada con un análisis Transformer para estudiar las trayectorias de escritura, formando así un sistema dual imagen-secuencia. El corpus se basa en un modelo preentrenado CASIA-HWDB2.2, recopilando datos a través de crowdsourcing provenientes de aprendices en 23 países, que tras ser preprocesados con OpenCV y anotados por expertos, resultan en un conjunto de datos que cubre 11,230 caracteres correspondientes a los niveles HSK1-6. El modelo utiliza un aprendizaje por transferencia en dos fases: primero se entrena una red ResNet-50 sobre el conjunto de datos estándar impreso CTD-HCC para extraer características, y luego se ajusta el mecanismo de atención sobre HCCD-2.0, logrando una precisión del 89.7% en el reconocimiento del orden del trazo (F1=0.872), significativamente superior a los métodos tradicionales ($p<0.01$). A través de pruebas A/B con 120 aprendices intermedios, se encontró que el grupo con retroalimentación en tiempo real basada en IA mostró ventajas significativas tanto en la tasa de retención memorística -aumento del 21.3%- como en la fluidez al escribir (reducción del tiempo por carácter en 18.5 segundos) en comparación con el grupo control ($p<0.05$). Los puntos innovadores incluyen: la creación del modelo cerrado cognición-acción-retroalimentación, un algoritmo personalizado para corrección basado en LSTM-GAN, el desarrollo de una interfaz multiplataforma HCAI-API, y el establecimiento de un sistema para medir la dificultad al escribir que incluye parámetros como complejidad topológica y otros 12 dimensiones. La investigación confirma que la inteligencia artificial puede aliviar eficazmente la fragmentación cognitiva entre forma-sonido-significado en los caracteres chinos, promoviendo así una transformación en la enseñanza del chino desde un enfoque basado en experiencias hacia uno impulsado por datos.

**Palabras clave:** IA, caracteres chinos, enseñanza, innovación

## Referencias

Gao, J. (2019). Investigación sobre el posicionamiento y la transformación del rol docente en el contexto de la IA. *Revista de la Escuela Técnica y Vocacional de Jincheng*, 12(4), 61-63. https://doi.org/10.3969/j.issn.1674-5078.2019.04.017

Hu, J. L. y Yin, Y. X. (2007). Aplicación de la tecnología de inteligencia artificial en la educación. *Conocimientos y tecnología informática*, 2(12), 1667-1668. https://doi.org/10.3969/j.issn.1009-3044.2007.12.096

Ma, N. (2022). Un nuevo camino para la enseñanza internacional del carácter chino en el contexto de la educación inteligente con IA. *Computadora Fujian*, 38(12), 118-123. https://doi.org/10.16707/j.cnki.fjpc.2022.12.025

Wei, S. Y. (2024). Investigación sobre las estrategias de construcción de caminos internacionales de enseñanza del carácter chino en la era de la IA. *Cultura del carácter chino*, (24), 75-77. https://doi.org/10.14014/j.cnki.cn11-2597/g2.2024.24.033

Yu, H. L. (2024). El cambio en el paradigma de enseñanza de lenguas extranjeras en colegios y universidades en la era inteligente. *Enseñanza e investigación de lenguas extranjeras*, 56(6), 913-923. https://doi.org/10.19923/j.cnki.fltr.2024.06.006

# Análisis del Discurso y Lingüística del Corpus para Enseñanza de Lenguas con Fines Sociales

*Antonio Jesús Tinedo Rodríguez (Universidad de Córdoba, Spain)*

La lingüística del corpus permite abordar el análisis del discurso aplicando métodos computacionales que facilitan el análisis de grandes cantidades de texto, encontrando patrones y especificidades de diferentes géneros discursivos y textuales. Estos textos no dejan ser el reflejo de la interacción entre la lengua, la sociedad y el pensamiento, por lo que supone una fuente muy rica de datos. Existen diferentes marcos teóricos en el ámbito del análisis del discurso, pero desde un punto de vista del análisis social, el Análisis Crítico del Discurso (Fairclough, 1992) supone un eje que permite explorar dinámicas de poder y desigualdades que, desde el prisma pedagógica, tienen un gran interés. El presente estudio parte del sustrato teórico que proporciona la intersección de la Lingüística del Corpus y el Análisis Crítico del Discurso, y cómo esta simbiosis permite diseñar tareas para el aprendizaje de lenguas. Para ello, se ha compilado un corpus denominado ECHR-Gen38 que está compuesto por los 38 fallos más relevantes del Tribunal Europeo de Derechos Humanos relacionados con el género. A través de *software* especializado se han analizado estos fallos y se han encontrado las temáticas recurrentes en lo que concierne al género. A partir de ellos se han diseñado un conjunto de tareas para que aprendices de lenguas puedan, a partir de material auténtico, no solo aprender cómo se utiliza la terminología especializada en contexto, sino que también esta propuesta cumpla con un doble fin pedagógico, facilitar la exposición a la lengua en contextos reales, y aplicar la mediación para reflexionar sobre desigualdades sociales, perpetuaciones de violencias y sus ecos en los diferentes géneros discursivos y textuales. Los hallazgos de este estudio, desde el punto de vista textual, reflejan que los textos jurídicos reflejan claramente múltiples situaciones de violencia contra las mujeres y que el alumnado de lenguas valora positivamente tanto la exposición a estos textos utilizando análisis como KWIC, y las reflexiones que estos suscitan. Las personas participantes han manifestado que el conjunto de tareas les resultó motivante, que les ayudó a mejorar su léxico y que tuvo un impacto positivo en su producción escrita. Asimismo, resaltaban la efectividad para abordar la conciencia de género desde una perspectiva intercultural puesto que los fallos que conformaban el corpus forman parte de diferentes países.

**Palabras clave:** Análisis Crítico del Discurso; Lingüística del Corpus; Enseñanza de Lenguas; Género

**Referencias**

Fairclough, N. (1992). *Discourse and social change*. Cambridge.

# Integrating parallel corpora into the EFL classroom: A practical case with English idioms and PaEnS

*Sidoní López Pérez (Universidad Internacional de La Rioja, Spain)*

Since the 1980s, corpus linguistics has experienced rapid growth. Although English language corpora continue to dominate research in this field, the development and utilization of corpora in other languages have significantly contributed to the expansion of studies based on corpus linguistics (McEnery and Xiao, 2007). As Álvarez Gil (2019) points out, corpus linguistics offers significant benefits to language learners, as the collected texts used in their studies usually serve as authentic examples of the target language. In this sense, corpus-based materials generally expose students to real-world language use while allowing them to engage with genuine linguistic productions in their learning process. Apart from monolingual corpora, parallel corpora have emerged as a central focus in non-English corpus linguistics, especially because of their crucial role in translation studies and contrastive linguistic research (McEnery and Xiao, 2007). Parallel corpora, defined as "corpora that contain a series of source texts aligned with their corresponding translations" (Malmkjaer, 1998, p. 539), can be used for various applications, including four main areas of focus: contrastive linguistics and translation studies, translation practice, lexicography, and the growing fields of foreign language and translation teaching (Doval and Sánchez, 2019).

However, the use of parallel corpora has been somewhat limited in the classroom. Different factors such as the belief that the language in corpora is too complex and difficult for learners, the perception that the tools for searching corpora are not user-friendly, and issues like a lack of awareness among instructors and inaccessible tools for searching parallel corpora contribute to this limitation (Brown, 2017; Doval and Sánchez, 2019). Despite these challenges, parallel corpora hold significant potential and can be used effectively in foreign language teaching with upper-intermediate and advanced students (Doval, 2018).

In this sense, the present study aims to present the potential use of the English/Spanish parallel corpus, PaEnS, in an English as a Foreign Language (EFL) classroom for students at an advanced level. To illustrate this approach, the study includes a brief practical case focusing on the teaching of some English idioms and their Spanish equivalents as found in the corpus. The reason for selecting idioms comes from the observation that learners often

struggle with these grammar structures, as it is generally impossible to guess what idioms mean from the words they contain (*Cambridge International Dictionary of Idioms*, 1998). To this effect, ten idioms will be selected according to *English Idioms in Use: Advanced* (2017), and the students will need to search for instances of these idioms in PaEnS and analyze their Spanish translations. After this, they will categorize the translations based on equivalence, paraphrasing, and omission (Baker, 1992). Finally, they will use ChatGPT to provide alternative Spanish translations, allowing for a comparison between human and AI-generated interpretations. By integrating PaEnS and ChatGPT into idiom instruction in the EFL classroom, students are expected to enhance their comprehension and competence in English idioms while simultaneously learning about and understanding their Spanish translations through data-driven analysis. At the same time, the findings are expected to contribute to increased awareness of how parallel corpora can support language learning, thereby helping bridge the gap between corpus research and pedagogical practice.

**Keywords:** corpus linguistics, parallel corpora, EFL, English idioms, PaEnS, ChatGPT.

## References

Álvarez Gil, F. J. (2019). Enseñanza de pragmática en lengua inglesa a nivel universitario a través del uso de metodología de corpus. *DEDiCA. Revista de Educação e Humanidades,* 15, 161-172. https://doi.org/10.30827/dreh.v0i15.8057

Baker, M. (1992). *In other words: A course book on translation.* Routledge.

Brown, M. (2017). Using parallel corpora for language learning. *Humanising Language Teaching*, *19*(3). Retrieved January 15, 2025, from https://www.researchgate.net/publication/328094165_Using_Parallel_Corpora_for_Language_Learning/download

*Cambridge International Dictionary of Idioms* (1998). Cambridge University Press.

Doval, I. (2018). Corpus paralelos en la enseñanza de lenguas extranjeras: un ejemplo de aplicación basado en el corpus PaGeS. *CLINA, 4*(2), 65–82. https://doi.org/10.20868/clina.2018.4.2.313

Doval, I., & Sánchez Nieto, M. T. (2019). Parallel corpora in focus: An account of current achievements and challenges. In I. Doval & M. T. Sánchez Nieto, M. T. (Eds.), *Parallel corpora for contrastive and translation studies: New*

*resources and applications* (pp. 1-15). John Benjamins Publishing Company. https://doi.org/10.1075/scl.90.01dov

Malmkjaer, K. (1998). Love thy neighbour: Will parallel corpora endear linguists totranslators? *Meta, 43*(4), 534–541. https://doi.org/10.7202/003545ar

McCarthy, M., & O'Dell, F. (2017). *English idioms in use: Advanced* (2nd ed.). Cambridge University Press.

McEnery, A. M., & Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to? In G. James, & G. Anderman (Eds.), *Incorporating corpora: The linguist and the translator* (pp. 18-31). Multilingual Matters.

# Evaluación del rendimiento de modelos de IA en la selección y corrección de los Complementos de Régimen Preposicional en ELE

*Wanyun Qi (Universidad de Valladolid, Spain; Universidad de Ciencia y Tecnología de Chongqing, China)*

Los Complementos de Régimen Preposicional (CRP) han sido objeto de estudio en la gramática española desde Alarcos Llorach (1966, 1990, 1994), quien los estableció como una categoría gramatical independiente dentro del predicado. Investigaciones recientes han analizado su distribución en español actual (Casanova Romero, 2021) y su enseñanza en ELE (Canales Muñoz, 2021), identificando dificultades en su adquisición debido a la fijación verbo-preposición y la interferencia de la lengua materna.

En este sentido, en el aprendizaje de ELE, el análisis de errores (Corder, 1967) ha sido clave para describir patrones recurrentes en el uso de los CRP, observándose similitudes en errores entre aprendientes de diferentes lenguas maternas (Alexopoulou, 2006).

El desarrollo de la inteligencia artificial (IA) ha abierto nuevas posibilidades en la enseñanza de ELE, particularmente en la retroalimentación automatizada. Investigaciones recientes han analizado su aplicación en la enseñanza de lenguas extranjeras (Fernández, 2024; Hernández, 2021; Ribes-Lafoz & Navarro Colorado, 2023), pero su efectividad en la selección y corrección de CRP aún no ha sido explorada en profundidad.

Este estudio evalúa el rendimiento de modelos de IA en la selección y corrección de los CRP en comparación con hablantes nativos y aprendientes de ELE. Se diseñó un experimento con tres grupos: hablantes nativos (grupo de control), aprendientes de ELE (grupo experimental 1) y modelos de IA (grupo experimental 2). En la primera fase, los participantes realizaron una prueba de selección de CRP para comparar la precisión en la elección de preposiciones y los patrones de error. En la segunda fase, se seleccionaron 25 errores representativos de los aprendientes y fueron corregidos por los modelos GPT-4o, Mini 4o y DeepSeek, analizando la precisión de las correcciones y la calidad de la retroalimentación proporcionada.

Los análisis preliminares indican que los modelos de IA presentan similitudes en la selección de preposiciones con los hablantes nativos en ciertos contextos, aunque también replican errores comunes en los aprendientes de

ELE. Se identificaron diferencias en la retroalimentación de los modelos, con algunos proporcionando respuestas correctas sin explicaciones detalladas y otros generando correcciones inconsistentes. Este estudio busca contribuir al análisis del potencial de la IA en la enseñanza de ELE y su posible integración en la enseñanza y evaluación de los CRP.

**Palabras clave:** Complementos de Régimen Preposicional, ELE, inteligencia artificial, modelos de lenguaje, Análisis de Error

## Referencias

Alarcos Llorach, E. (1966). Verbo transitivo, verbo intransitivo y estructura del predicado. *Archivum: Revista de la Facultad de Filosofía y Letras*(16), 5-17.

Alarcos Llorach, E. (1990). *La noción de suplemento*. Consejería de Educación, Cultura y Deportes.

Alarcos Llorach, E. (1994). Gramática de la lengua española. *(No Title)*.

Alexopoulou, A. (2006). Los criterios descriptivo y etiológico en la clasificación de los errores del hablante no nativo: una nueva perspectiva. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras*(5), 17-36. https://dialnet.unirioja.es/servlet/articulo?codigo=1709311

Canales Muñoz, L. A. (2021). El complemento de régimen preposicional en la enseñanza de español como lengua extranjera. *E-eleando*, *n 21*, 1-107. https://doi.org/https://ebuah.uah.es/dspace/handle/10017/49133

Cárdenas, J. (2023). Inteligencia artificial, investigación y revisión por pares: escenarios futuros y estrategias de acción [Derechos de autor 2023 Julián Cárdenas]. *Revista Española de Sociología*, *32*(4), a184-a184. https://recyt.fecyt.es/index.php/res/article/view/101519

Casanova Romero, V. (2021). *El complemento de régimen verbal: construcción y distribución en español actual* [Tesis doctoral, Université de Montréal].

Corder, S. P. (1967). La importancia de los errores del que aprende una lenua segunda (traducido de THE SIGNIFICANCE OF LEARNER'S ERRORS). In *La adquisición de las lenguas extranjeras* (1992 ed., pp. 31-40). Visor. https://eric.ed.gov/?id=ED019903

Fernández, M. G. (2024). Inteligencia artificial y léxico: Una propuesta con ChatGPT para niveles B1 y B2 en la enseñanza del español como lengua extranjera [Derechos de autor 2024 María García Fernández]. *RILEX*.

*Revista sobre investigaciones léxicas*, e9111-e9111. https://revistaselectronicas.ujaen.es/index.php/RILEX/article/view/9111

Hernández, J. C. E. (2021). La Inteligencia Artificial y la Enseñanza de lenguas: una aproximación al tema [Derechos de autor 2021]. *Decires*, *21*(25), 29-44. https://decires.cepe.unam.mx/index.php/decires/article/view/3

Ribes-Lafoz, M., & Navarro Colorado, B. (2023). Aprovechamiento de ChatGPT en la enseñanza de lengua extranjera en educación superior. In. Octaedro. http://hdl.handle.net/10045/139198

# Tryna Learn English with Literary Texts

*Michael Lang*

Data-driven learning (DDL) is an approach to the language learning in which students interact directly with corpus data, either through an online interface or through activities that have been tailored to their needs. Often referred to as a form of *bottom-up* or *inductive learning*, with DDL students become 'language detectives' and take an active role in the learning process.

It is through this framework that we consider the corpus-based exploration of literary texts by foreign language students. We argue that the use of parallel literary texts allows for a richer understanding of language variation as well as prosody compared to other text domains.

The PaEns[1] corpus contains approximately 100 English-language literary texts alongside their Spanish translations. By analyzing instances of the verb *try* in the corpus data, particulary within fictional dialogue, we consider what insights literary texts can provide students when it comes to such difficults aspects of English as variation and prosody.

For variation, we examine the usage of *try to* vs *try and* in the data. While students are taught the standard *try to* (e.g. *I'll try to get in there with that*) to express an attempt to carry out an action, they are often less aware of the variant *try and* (e.g. *I'll try and get in there with that*). This section explores how students can discover the possibilites and limitations of these two variants.

As for prosody, literary authors often strive to represent the way speakers pronounce words or combinations of words. This provides a wealth of opportunities for langauge students as the contrast between the familiar standard spelling and the more phonetic representation of speech in the text can prove incredibly informative for students. We specifically look at the usage of *tryna* (a blended form of *trying to*) and what that can tell students about vowel reduction in English prosody.

**Keywords:** Data-driven Learning, ESL instruction, prosody, literary texts, Second Language Acquisition

## References

Boulton, A. (2021). Research in data-driven learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond the concordance: Corpora in language education* (pp. 9-34). Amsterdam: John Benjamins.

Lind, Å. (1983). The variant forms *try and/try to*. *English Studies*, 64(6), 550–563. https://doi.org/10.1080/00138388308598291

Vyatkina, N. (2020). Corpora as open educational resources for language teaching. *Foreign Language Annals*, 53(2), 359–370. https://doi.org/10.1111/flan.12464

# 6. Traducción multilingüe y redacción profesional / Multilingual translation and professional writing

# The Role of Multilingual Translation in Enhancing Professional Writing Skills Across Cultures: A Study on Corporate Communication and Global Market Strategies

*Kamal Khazal Mohammd (Universidad de Valladolid, Spain; Imam Al Kadhum College, Iraq)*

*Narjis Audah Rashk (University of Misan, Iraq)*

This research investigates multilingual translation with the view of professional writing skills especially in corporate and global marketing areas and marketing strategies. Due to globalization and all of that, there has to be an effort by businesses communicating across cultural and lingual borders to be clear, consistent, and engaging in their messaging. This study discusses multilingual translation practices contributing to the improvement of professional writing regarding Coca-Cola as a company, which is a very relevant multinational corporate entity with operations in many markets worldwide. The case study analyzes Coca-Cola's marketing communications in English, Arabic, and Spanish, with a focus on how translation and localization of marketing communication materials, advertisements and internal corporate communications affect the effectiveness of their written contents from one culture to another. The population sample of this study is comprised of 40 professional writers, translators and marketers who work in the communication and branding departments of Coca-Cola collaboration through the translation processes and also feedback from regional markets in the Middle East and Latin America. The mixed method comprises surveys, in-depth interviews of key communication staff, and content analysis of translated campaign and internal documents to determine how multilingual translation practices affect corporate message tone, cultural appropriateness and engagement. Also, some professional writing skills are required in adapting them to local cultural variance while keeping the identity intact on a global level.

The results show that multilingual translation leverages professional writing in understanding messages with specific audiences without cultural interference. More precisely, the present investigation emphasizes the criticality of sensitivity with which marketing materials are construed and the capability of translators to bridge the linguistic and cultural gaps in understanding between the various languages into which messages are to be rendered. This research illustrates how multilingual translation develops professional-writing skills in light of global market strategies and offers recommendations for multinational

corporations on adopting multilingual communication strategies into their operations. It actually concludes that a scope of translation with cultural informativity will not only strengthen brand consistency but also improve market engagement and loyalty in varying regions of the global marketplace.

# *Est-ce que* our particles are the same *ma?* A corpus-based translation study of question particles in Mandarin and French

*Hung-Hsin Hsu (Université catholique de Louvain, Belgium; National Chengchi University, Taiwan)*

Typologically, questions can be marked phonetically (e.g., intonation) or morpho-syntactically (e.g., reordering, particle, tag) (Siemund, 2001; Hölzl, 2018). Question particles are used in Mandarin and French. In Mandarin, several question particles are available, such as 嗎ma, 吧ba, 啊a, 呢ne (Fang & Hengeveld, 2022, p. 879), as in (1) and (2).

(1) 你是學生**嗎/吧/啊**?
Nǐ shì xuéshēng **ma/ba/a**?
'Are you a student?'

(2) 誰是老師**啊/呢**?
Shéi shì lǎoshī **a/ne**?
'Who is the teacher?'

Ma and ba can only be used in polar questions (also known as yes-no questions), while ne only occurs in wh- and disjunctive questions. The a particle is allowed in any question type.

In French, est-ce que is considered as a question particle (Druetta, 2018), as in

(3) Est-ce qu'il vient ?
'Is he coming?'

The *est-ce que* particle occurs in any question type. However, two other particles, *ça* and *donc*, only appear in *wh*-questions, as in (4) and (5).

(4) Qui **ça** les autres ?
Who are the others?' (Druetta, 2018, p. 34, my translation)

(5) Où **donc**, là-haut ?
'Where up-there?' (Smirnova & Abeillé, 2021, p. 259)

Given this common ground in both languages, I aim to examine the cross-linguistic equivalence of question particles in Mandarin and French, drawing on parallel data from a self-compiled bidirectional Mandarin and French corpus of contemporary novels. My main research question is: How are particles translated across Mandarin and French?

Previous work shows that particles in Asian languages often fulfil functions that are conveyed by intonation in European languages (Hancil et al., 2015, p. 18). Also, question particles in Mandarin can be translated primarily by intonation (i.e., using declarative word order), followed by inversion, and tags (Chao, 1968, p. 804). Accordingly, the following hypothesis can be formulated: question particles in Mandarin are translated into French primarily by intonation (i.e., using declarative word order), and vice versa.

Parallel texts were manually aligned using Intertext (Vondřička, 2014) and uploaded to Sketch Engine (Kilgarriff et al., 2014). Questions were automatically extracted using the advanced Concordance function (search on the character: ?). The number of questions in each source language (SL) is 3,921 in Mandarin and 4,224 in French. All SL instances of questions and their translations in the target language (TL) were manually coded for syntactic structures (e.g., inversion in French, A-not-A structure in Mandarin), and the presence of a particle, such as *ma*, *ba*, *a*, *ne* in Mandarin, and *est-ce que*, *ça*, *donc* in French.

Preliminary results indicate that the number of particles decreases in translations from Mandarin to French (from 961 to 416), while the number of particles in Mandarin translations is much higher than in the French source texts (from 254 to 1,442). In my talk, I will discuss the recurrent patterns of cross-linguistic equivalences uncovered in the parallel corpus and explore the theoretical implications of the findings for the study of questions in contrastive linguistics.

**Keywords:** contrastive study, parallel corpus, question particle, Mandarin, French

**References**

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Universtiy of California Press.

Druetta, R. (2018). Syntaxe de l'interrogation en français et clivage écrit-oral: Une description impossible ? In M.-J. Béguelin, A. Coveney, & A. Guryev (Eds.), *L'interrogative en français* (Vol. 124, pp. 19–50). Peter Lang CH. https://doi.org/10.3726/b13079

Fang, H., & Hengeveld, K. (2022). Sentence-Final Particles in Mandarin. *Studia Linguistica*, *76*(3), 873–913. https://doi.org/10.1111/stul.12198

Hancil, S., Post, M., & Haselow, A. (2015). 1. Introduction: Final particles from a typological perspective. In S. Hancil, A. Haselow, & M. Post (Eds.), *Final*

*Particles* (pp. 3–36). DE GRUYTER. https://doi.org/10.1515/9783110375572-001

Hölzl, A. (2018). *A Typology Of Questions In Northeast Asia And Beyond: An Ecological Perspective*. Zenodo. https://doi.org/10.5281/ZENODO.1344467

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Siemund, P. (2001). Interrogative Constructions. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher and Wolfgang Raible (Ed.), *Language typology and language universals* (pp. 1010–1028). Walter de Gruyter.

Smirnova, A., & Abeillé, A. (2021). Question particles ça and donc in French: A corpus study. *Linguistic Research*, *38*(2), 239–269. https://doi.org/DOI:10.17250/khisli.38.2.202106.003

Vondřička, P. (2014). Aligning parallel texts with InterText. In N. C. (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1875–1879). European Language Resources Association (ELRA).

# Un estudio comparativo de la semántica en la traducción entre chino y español

*Wanruo Luo (Universidad de Salamanca, Spain)*

Este trabajo se centra en la relación interactiva entre los símbolos lingüísticos y la lógica cultural profunda en la comunicación intercultural. El objetivo es revelar, a través de un análisis comparativo sistemático, las diferencias semánticas en las expresiones metafóricas, conceptos folclóricos, referencias históricas y valores sociales entre el chino y el español, así como explorar cómo lograr una transmisión cultural equivalente mediante estrategias de traducción dinámicas. La investigación se basa en las teorías de traducción semántica y comunicativa de Newmark, combinando la perspectiva de traducción cultural de Bassnett y el concepto de filtro cultural propuesto por Baker. Construimos un marco analítico dual que considera tanto la forma lingüística como la función cultural, centrándonos en tres problemas centrales que aparecen con frecuencia en la traducción entre chino y español: vacíos léxicos, generalización semántica y desplazamiento de imágenes. Desde el punto de vista metodológico, el estudio adopta un enfoque impulsado por un corpus mixto, seleccionando cuatro tipos de textos especializados para construir un corpus paralelo: clásicos literarios, documentos legales, guías médicas y subtítulos cinematográficos. Utilizamos técnicas de procesamiento del lenguaje natural para extraer unidades semánticas culturales, combinadas con anotaciones manuales para identificar diferencias en los mapeos conceptuales metafóricos -como la comparación entre el sistema simbólico del dragón en chino y el águila en español-. Además, cuantificamos la frecuencia de aplicación de técnicas como la traducción literal, sustitución y compensación a través de una matriz de estrategias traductoras. Los hallazgos indican que las diferencias culturales y semánticas entre el chino y el español pueden clasificarse en tres categorías: percepción temporal-espacial -por ejemplo, desajuste en las imágenes climáticas del viento del este/viento del oeste-, valores éticos -como la desigualdad categórica entre filialidad y honor familiar- y cultura material -como las diferencias simbólicas festivas entre zongzi y turrón-. El tratamiento traductológico debemos elegir estrategias basadas en las diferencias funcionales del texto: la traducción literaria se centra en la intertextualidad cultural dentro de la reconstrucción de imágenes -como la transferencia creativa de metáforas rurales en las novelas de Mo Yan-, mientras que la traducción técnica enfatiza la estandarización terminológica y la contextualización -como la conversión explicativa del término médico chino

shanghuo dentro del sistema médico hispano-. Por otro lado, la traducción audiovisual debe equilibrar expresiones coloquiales con compensaciones culturales -como las técnicas de anotación instantánea para modismos o referencias culturales en los subtítulos-. Este estudio propone establecer una base de datos sobre semántica cultural chino-española y un modelo dinámico para evaluar estrategias, validando empíricamente el impacto de diferentes enfoques sobre la aceptación del texto traducido. Finalmente, buscamos desarrollar una guía para la traducción intercultural que contemple tanto normas académicas como necesidades industriales, proporcionando un referente teórico para optimizar la adaptabilidad cultural en sistemas de traducción automática y promoviendo una transformación en la enseñanza traductora desde una superficie lingüística hacia el desarrollo profundo de habilidades cognitivas culturales.

**Palabras clave:** Multilingüe, traducción, comparación, semántica

## Referencias

He, R. N. (2024). Un estudio sobre las características lingüísticas y estrategias de traducción de refranes chinos y españoles. *Lingüística moderna*, 12(10), 785-791. https://doi.org/10.12677/ml.2024.1210955

Jin, R. (2022). *Un estudio comparativo de los adverbios de postura en el discurso informativo chino y español, tomando como ejemplos "Te Bie" y "especialmente"* [Trabajo de Fin de Máster]. Universidad de Estudios Internacionales de Xi'an, Xi'an. Recuperado de https://doi.org/10.27815/d.cnki.gxawd.2022.000338 [Fecha de consulta: 19/01/2025].

Mombelli, D. (2019). La metodología comparatista en los estudios literarios. *Revista Española de Educación Comparada*, (34), 97–117. https://doi.org/10.5944/reec.34.2019.24379

Nie, Q. (2020). *Diferencias entre chino y español y la huida de la interpretación consecutiva chino-español* [Trabajo de Fin de Máster]. Universidad de Estudios Extranjeros de Pekín, Pekín. Recuperado de https://doi.org/10.26962/d.cnki.gbjwu.2020.000052 [Fecha de consulta: 20/01/2025].

Tong, Y. X. (2015). La influencia y contramedidas del pensamiento chino en la traducción en la etapa básica del español. *Educación de calidad occidental*, 1(2), 35-36. https://doi.org/10.16681/j.cnki.wcqe.2015.02.008

# Sugerencia automática de maridajes: un corpus paralelo para la traducción y la recomendación de combinaciones enogastronómicas

*María Pascual Cabrerizo (Universidad de Valladolid, Spain)*

La gastronomía es un campo en el que las diferencias lingüísticas y culturales impactan directamente en la aceptabilidad de los textos traducidos o redactados para un público extranjero y uno de los elementos que más dudas puede presentar al traductor/redactor es el maridaje de recetas y vinos (Romero, 2020). La investigación que aquí presentamos surge con el objetivo principal de ayudar a estos profesionales de la lengua a resolver tales dificultades, bien con datos que contribuyan a la toma de decisiones de traducción, bien con un generador automático de sugerencias que puede complementar otras herramientas digitales generativas a disposición de los comunicadores plurilingües de los sectores vitivinícola y gastronómico, como los asistente de escritura derivados del proyecto ACTRES (Moreno Pérez y López Arroyo, 2021; Ortego Antón, 2024) .

El presente trabajo plantea la creación y utilización de un corpus paralelo especializado en maridajes de vino y comidas, con el objetivo de analizar sus aplicaciones en la traducción y en sistemas de recomendación basados en inteligencia artificial (IA). Para ello, hemos diseñado un corpus bilingüe (español-inglés) compuesto por descripciones de maridajes extraídas de fichas de cata de vinos españoles y estadounidenses y literatura gastronómica reciente, estructuradas mediante anotación lingüística y semántica (basadas en tipos de vino, variedades de uva, regiones, perfiles de sabor, tipos de platos, elaboraciones, ingredientes y justificación del maridaje). En realidad, se trata de un corpus compuesto de un subcorpus paralelo, útil para la comparación de originales y traducciones, y otro comparable, más útil para la generación automática de recomendaciones y la predicción de los maridajes que podrían ser más aceptables en una u otra cultura.

En el momento de redactar esta propuesta, el corpus ya se considera representativo según los parámetros de la herramienta ReCor (Corpas Pastor y Seghiri Domínguez, 2010), pero nuestra intención es seguir nutriéndolo en los próximos meses. Los componentes esenciales del maridaje (tipo de vino, características organolépticas, plato recomendado e ingredientes principales) se están etiquetando automáticamente mediante técnicas de procesamiento del lenguaje natural (Bird, Klein y Loper, 2009), lo que permitirá entrenar

modelos de traducción automática neuronal en la siguiente fase de la investigación y nos ayuda a analizar patrones terminológicos y discursivos en el ámbito enogastronómico. Todas las muestras se están almacenando por separado y sin marcas en archivos de texto simple, compatibles con herramientas como Sketch Engine, y en versión etiquetada en un único archivo CSV listo para su uso en proyectos de IA.

Desde el punto de vista de la traducción, el análisis de las muestras tal cual se han recogido permite estudiar la variación léxica y la transferencia cultural en las recomendaciones enogastronómicas. Por otro lado, las muestras tratadas permiten explorar aplicaciones del corpus en la generación automática de maridajes mediante modelos de IA. En particular, estamos trabajando en el desarrollo de *pAIring*, un prototipo de aplicación basada en corpus que pretende ser de utilidad para traductores y redactores especializados, así como para profesionales de la restauración y consumidores. Este proyecto se encuentra todavía en una fase embrionaria en la que estamos probando modelos sencillos de Machine Learning (Random Forest), pero aspira al uso de arquitecturas de IA que vayan más allá de la clasificación y permitan entender realmente las relaciones semánticas en los maridajes (embeddings) para llegar a un modelo generativo final (transformers). El resultado esperado es que el modelo entrenado con este corpus mejore la coherencia y naturalidad de las traducciones en textos enogastronómicos, en comparación con sistemas generales de traducción automática. Además, la integración de información semántica podría facilitar la generación de recomendaciones de maridaje personalizadas a partir de las preferencias del usuario.

El trabajo muestra el potencial de los corpus especializados en una combinación de lingüística computacional, traducción automática y modelos de IA en el ámbito enogastronómico, pero todavía estamos lejos de poder comprobar su verdadera utilidad, para lo que habrá que iniciar una nueva investigación basada en encuestas cuando el corpus haya crecido significativamente y *pAIring* esté desarrolado a un nivel de funcionalidad que permita la prueba con usuarios reales. Como otra futura línea de investigación y acción, se contempla la ampliación del corpus hacia otros pares de lenguas.

**Palabras clave:** corpus, maridajes, generación automática, procesamiento del lenguaje natural, traducción enogastronómica

## Referencias

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

Corpas Pastor, G. y Seghiri Domínguez, M. (2010). Size Matters: A Quantitative Approach to Corpus Representativeness en R. Rabadán (Ed.) *Lengua, traducción, recepción. En honor de Julio César Santoyo/ Language, translation, reception. To honor Julio César Santoyo* (pp. 112-146). Secretar.

Moreno Pérez, L. y López Arroyo, B. (2021). Atypical Corpus-Based Tools to the Rescue: How a Writing Generator Can Help Translators Adapt to the Demands of the Market. *MonTI 13*(1), 251-279.

Ortego Antón, M. T. (2024). Designing a corpus-based writing aid tool for the agrifood industry: The case of TorreznoTRAD en F. Alonso Almeida (ed.), *Insights in (inter) cultural and cross-cultural communication* (pp. 51-66). Tirant lo Blanch.

Romero, L. (2020). La traducción gastronómica: un estudio sobre los problemas de traducción en las recetas en A. M. López Márquez y F. Molina Castillo (coord.) *Italiano y español. Estudios de traducción, lingüística contrastiva y didáctica* (pp. 253-269). Peter Lang.