DISCURSO ACADÉMICO, TRADUCCIÓN AUTOMÁTICA Y EVALUACIÓN DE LA CALIDAD: COMPARACIÓN ENTRE SISTEMAS

Eugenia Sainz
Università Ca' Foscari Venezia
Venecia, Italia

Antonella Bove
Università Ca' Foscari Venezia
Venecia, Italia

Resumen

El presente artículo quiere contribuir a la investigación sobre la evaluación de la calidad de la traducción automática con un estudio enfocado en la traducción del discurso académico del italiano al español y viceversa. Partiendo de la teoría de la pertinencia (Sperber y Wilson, 1986), el estudio compara la eficacia de los sistemas neuronales (en concreto, DeepL y Google Translate) con los generativos (ChatGPT-4o y Deepseek) a través de un análisis cuantitativo y cualitativo basado en la métrica MQM. Se trabaja a partir de un corpus de longitud media, constituido por diez resúmenes en italiano y diez en español. Los resultados validan las hipótesis de partida. Primero, los sistemas generativos demuestran un mejor rendimiento que los neuronales, con un número menor de errores graves. Segundo, la mayoría de los errores afecta a la reconstrucción de la explicatura y, concretamente, a las categorías de *Terminología* y *Precisión*. Tercero, se advierte un alto nivel de variación estilística, en el que los sistemas generativos destacan por su capacidad de reformulación y mejora del texto en términos de claridad. Cuarto, el rendimiento de los sistemas resulta mejor en la dirección italiano-español. Finalmente, se confirma la centralidad del traductor humano como garante de la calidad última del texto.

Palabras clave: evaluación de calidad de la traducción automática, métrica MQM, discurso académico español<>italiano, análisis de errores, estilo.

ACADEMIC DISCOURSE, MACHINE TRANSLATION AND QUALITY ASSESSMENT: A CROSS-SYSTEM COMPARISON

Abstract

This paper contributes to research on machine translation quality assessment through a study focused on the translation of academic discourse between Italian and Spanish. Drawing on the relevance theory (Sperber and Wilson, 1986), the study compares two neural machine translation systems (NMT) (DeepL and Google Translate) with two LLMs (ChatGPT-4o and Deepseek), through both quantitative and qualitative analysis based on the MQM framework. The analysis is conducted on a medium-sized corpus consisting of ten abstracts in Italian and ten abstracts in Spanish. The results confirm the initial hypotheses. First, LLMs exhibit superior performance compared to NMT systems, producing fewer serious errors. Second, most errors affect the reconstruction of explicature, specifically within the categories of terminology and accuracy. Third, a high degree of stylistic variation is observed, with LLMs standing out for their ability to rephrase and enhance texts in terms of clarity. Fourth, system performance proves to be better when translating from Italian into Spanish. Ultimately, the human translator is reaffirmed as the key agent responsible for ensuring the overall textual quality.

Keywords: machine translation quality assessment, MQM framework, academic discourse Spanish<>Italian, error analysis, style.

DOI: https://doi.org/10.25145/j.refiull.2026.52.14
Revista de Filología, 52; junio 2026, pp. 379-416; ISSN: e-2530-8548

1. INTRODUCCIÓN

La aplicación de la inteligencia artificial (IA) a la práctica de la traducción ha cambiado radicalmente la actividad profesional y ha abierto un nuevo campo de estudio de alto impacto e interés social (Kornacki *et al.*, 2024)[1]. En este nuevo paradigma digital, el traductor se convierte en un supervisor (pre y poseditor, transcreador) altamente especializado que combina habilidades y conocimientos humanos y específicamente lingüísticos con saberes y competencias tecnológicas avanzadas para optimizar el resultado final (Lozano Zahonero, 2021; Cid-Leal *et al.*, 2019; Torrejón y Rico, 2012). DePalma (2017) acuña el concepto de *traducción aumentada* y de *traductor aumentado*, evocando de este modo la simbiosis del ser humano con las nuevas herramientas tecnológicas. El traductor se coloca en el centro del proceso de aprendizaje de la máquina y se confirma como figura insustituible tanto en la fase de producción como en la de evaluación de la calidad del producto:

> In 2017 we isolated «augmented translation», a machine-driven but human-centric approach in which linguists work directly with MT and an array of other technologies that support them, but that leaves them in charge. Based on the innovation of adaptive neural MT, it allowed humans to participate in the real-time, on-line training of neural engines. Rather than position the professional translator in the reactive position of cleaning up after a machine, it put them right in the middle of the training exercise. (DePalma, 2021)

La traducción automática es un campo del procesamiento del lenguaje natural (PLN) que sorprende por el desarrollo vertiginoso experimentado en las últimas décadas[2]. La revolución llega en el 2013 (Kalchbrenner y Blunsom, 2013), con un cambio de paradigma que abandona el enfoque de la traducción automática estadística dominante hasta el momento y aplica una nueva técnica de aprendizaje profundo o *deep learning* denominada metafóricamente *red neuronal* (Holdsworth y Scapicchio, 2024; Yu, 2024). La red neuronal procesa los datos de entrada (*input*) en varias capas de *neuronas* artificiales que capturan matemáticamente, a través de representaciones vectoriales, las relaciones semánticas de las palabras en su cotexto o contorno verbal, tanto en la frase origen que está siendo procesada como en la frase

---

[1] El presente artículo es fruto de la estrecha colaboración de las autoras, que han trabajado conjuntamente en cada una de sus partes. No obstante, las secciones 1, 2, 4 y 5.1. son atribuibles a Antonella Bove y las secciones 3, 5.2, 5.3. y 6, a Eugenia Sainz. El artículo se encuadra en el proyecto de excelencia 2023-2027 del *Dipartimento di Studi Linguistici e Culturali Comparati* (DSLCC) de la Università Ca' Foscari Venezia, financiado por el Ministero dell'Università e della Ricerca. El proyecto se centra en los procesos de adaptación literaria, cultural y lingüística a la luz de la noción / marco antropológico de adaptación / adaptabilidad. Agradecemos a las personas revisoras sus valiosos comentarios, correcciones y sugerencias, tanto de forma como de contenido, que han contribuido de manera significativa a la mejora de este artículo.

[2] Para una historia reciente de los distintos sistemas de traducción y de las características de su estructura, véase Sharma *et al.* (2023) y Specia y Wilks (2016). De enfoque humanista, Mitchell (2022) y Cristianini (2024).

meta que está siendo producida. El resultado del algoritmo genera una salida (*output*) que, en nuestro caso, es una traducción (Koehn, 2020, p. 11; Forcada, 2017, p. 293).

El modelo se mejora en el 2014, con la incorporación de un mecanismo de atención que opera a escala global y local, y se perfecciona ulteriormente en el 2017, con un modelo de *deep learning* avanzado e intensivo de autoatención (*self-attention*) y multiatención (*multi-headed attention*) conocido como *Transformer* (Vaswani, 2017). Con la llegada y difusión a partir del 2020 de los modelos de lenguaje a gran escala (o LLM), la arquitectura *Transformer* se convierte en el estándar dominante (Sharma *et al.*, 2023, pp. 13-21; Alammar, 2018)[3] e introduce una nueva revolución en el ámbito de la traducción automática (Brown, 2020, p. 14; Hendy *et al.*, 2023). El diseño del *prompt* y la posibilidad de dialogar con la máquina vuelven dinámica la actividad de posedición y cambian radicalmente el flujo de trabajo, lo cual exige nuevas competencias técnicas al traductor.

La traducción automática neuronal (TAN) y los LLM se basan, como decíamos, en una arquitectura tipo *Transformer,* pero presentan, con todo, diferencias significativas en su estructura, finalidad y modalidad de entrenamiento (Hendy *et al.*, 2023; Balashov, 2025, p. 10). Por lo que se refiere a su estructura, los modelos más tradicionales de TAN emplean una estructura codificador-descodificador (*Encoder-Decoder*): el codificador procesa una oración de entrada (por ejemplo, en español) y el descodificador genera, a partir de ella, una nueva oración en el idioma de salida (por ejemplo, en italiano). Por su parte, los LLM de tipo GPT emplean una estructura de solo descodificador (*decoder-only*). El sistema recibe el *prompt,* es decir, una instrucción en lenguaje natural que indica al modelo la tarea que tiene que hacer, acompañada de informaciones relevantes para llevarla a cabo de la mejor manera posible. El sistema recibe conjuntamente el *prompt* y la entrada (el texto que debe traducir) y, a partir de ahí, sin interrumpir el flujo, genera la traducción. El diseño de *prompt* para la creación de instrucciones precisas, claras y, en consecuencia, efectivas, comparte los supuestos del *lenguaje claro* (Da Cunha, 2024) y constituye actualmente un tema emergente de investigación y una de las competencias más relevantes del traductor.

Por lo que atañe a la modalidad y finalidad de entrenamiento y aprendizaje, los TAN (como Google Translate o DeepL) son sistemas únicamente entrenados y desarrollados para traducir. Aprenden a traducir de forma explícita, después de un entrenamiento específico sobre millones de frases paralelas que proceden de corpus multilingües cuidadosamente elaborados para ese propósito (Forcada, 2017; Brown *et al.*, 2020). En cambio, los LLM (como ChatGPT y Deepseek) se entrenan principalmente a partir de corpus monolingües de grandes dimensiones no revisados por humanos y concebidos, originariamente, con el único objetivo de aprender a predecir «una continuación plausible de un texto» (Gómez-Rodríguez, 2025, p. 77). Durante el proceso de entrenamiento, se observó, sin embargo, con sorpresa, que los modelos

[3] Hay programas informáticos que se basan inicialmente en el modelo estadístico y que pasan después al neuronal, como Google Translate y Microsoft Translator; otros, en cambio, nacen ya como neuronales (es el caso de DeepL y Amazon Translate).

no se limitaban a predecir la palabra siguiente, sino que, antes incluso de la fase de *fine-tuning* o refinamiento, desarrollaban habilidades no previstas, como la habilidad de responder a preguntas, de realizar operaciones aritméticas o de traducir. Fue un hallazgo inesperado: los modelos se revelaban «capaces de aprender sin supervisión explícita» (Cristianini, 2024, pp. 41-44). Eran capaces de traducir porque habían desarrollado una interlengua, es decir, habían interiorizado y aprendido regularidades interlingüísticas y correspondencias semánticas a partir de los patrones estadísticos extraídos de los datos de entrenamiento (Balashov, 2025). Es lo que Briakou *et al.* (2023) denominan *bilingüismo incidental*. De ahí que la capacidad traductora de los LLM pueda considerarse una *habilidad emergente*, esto es, «una capacidad que no se perseguía desarrollar directamente durante el entrenamiento, pero que el modelo aprendió igualmente» (Pantcheva, 2025; la traducción es nuestra)[4].

Debido, precisamente, a la reciente difusión de los LLM en la práctica traductora, es todavía escasa la investigación orientada a la evaluación comparativa de su grado de eficacia frente a los sistemas de TAN, en particular en la combinación lingüística español<>italiano[5]. El presente estudio quiere contribuir a colmar este vacío, focalizando la atención en un ámbito de especialidad concreto –el académico– a partir de un corpus de creación propia constituido por diez resúmenes en español y diez en italiano.

2. OBJETIVOS E HIPÓTESIS DE INVESTIGACIÓN

La investigación se plantea dos objetivos principales: en primer lugar, evaluar empírica y comparativamente la calidad de los resultados y el grado de eficacia de dos sistemas TAN (en concreto, DeepL y Google Translate) frente a dos LLM (en concreto, ChatGPT y Deepseek); y, en segundo lugar, identificar, clasificar y describir cualitativamente los tipos de errores generados, así como identificar fenómenos y patrones específicos de traducción automática que no se verifican en la traducción humana. Afrontamos el análisis intensivo y sistemático del corpus para explorar tendencias, ofrecer descripciones cualitativamente precisas y generar hipótesis que esperamos puedan servir como referencia o punto de partida para futuras investigaciones sobre otros ámbitos de especialidad o con corpus de mayor alcance.

[4] La intensa investigación de los últimos años ha hecho posible la aplicación de técnicas de refinamiento de los LLM para la traducción (Pantcheva, 2025), así como la aparición en 2024 de los primeros LLM específicamente creados y entrenados para traducir, como Lara y Widn.AI (desarrolladas respectivamente por Translated y Unbabel).

[5] La mayoría de los estudios realizados en el ámbito de la evaluación de la calidad de la traducción automática se centran en el inglés y, en menor medida, en el español como lengua de partida o de llegada (Hendy *et al.*, 2023; Kocmi *et al.*, 2024; Riina *et al.*, 2024). El par italiano<>español en concreto es representado de forma muy marginal (Minervini, 2021, 2023); de ahí el interés de la presente investigación.

El objetivo no es de generalización estadística, sino descriptivo, cualitativo y exploratorio. Partimos de los siguientes supuestos:

(i) Los sistemas generativos tendrán un mejor rendimiento que los neuronales;
(ii) Dada la tipología y el dominio textual, el mayor número de errores se concentrará en la reconstrucción de la explicatura de nivel inferior[6];
(iii) Los sistemas generativos serán capaces de activar procesos de elaboración discursiva más complejos que los neuronales con impacto en el estilo global del texto en términos de claridad y de accesibilidad;
(iv) La mayor presencia y la mejor posición relativa del español en la red y como lengua meta de traducción podrían reflejarse en un menor número de errores en la dirección IT>ES[7];
(v) En todos los casos, será necesaria la intervención del traductor humano, en cuanto garante de la calidad última del texto.

3. MARCO TEÓRICO

La investigación en traducción automática se encuadra en un marco multidisciplinar en el que se cruzan el PLN y la teoría de la traducción, que se sustenta, a su vez, en los distintos enfoques de la lingüística, los estudios culturales y la comunicación. Para nuestro estudio, el modelo ostensivo-inferencial propuesto desde la teoría de la pertinencia (Sperber & Wilson, 1986; Wilson & Sperber, 2004, 2012) resulta particularmente útil para dar cuenta de la gravedad de los errores y para explicar los problemas concretos que pueden llegar a verificarse durante el proceso de traducción de textos académicos. Asumimos, por tanto, que el texto traducido es un tipo particular de *metarrepresentación*, es decir, una representación que representa otra previa

[6] La teoría de la pertinencia introduce la noción de *explicatura* para evitar los problemas que planteaba la noción griceana del significado proposicional. A diferencia de lo supuesto por Grice (1975), la recuperación de la proposición no se obtiene solo por descodificación, sino que requiere también una importante dosis de inferencia. Como explican Wilson y Sperber (2004, pp. 614-615):

> Although the decoded logical form of an utterance is an important clue to the speaker's intentions, it is now increasingly recognized that even the explicit content of an utterance may go well beyond what is linguistically encoded. By 'explicitly communicated content' (or EXPLICATURE), we mean a communicated proposition recovered by a combination of decoding and inference, which provides a premise for the derivation of contextual implications and other cognitive effects.

La comprensión humana del significado explícitamente comunicado se obtiene, pues, «... via decoding, disambiguation, reference resolution, and other pragmatic enrichment processes» (Wilson y Sperber, 2004, p. 615). Se distingue entre *explicatura de nivel inferior,* o proposición propiamente dicha susceptible de ser evaluada en términos de valor de verdad, y *explicatura de nivel superior*, relacionada con la fuerza ilocutiva y la actitud.

[7] Véanse Instituto Cervantes (2024, pp. 81-82) y Rehm (2025, diapositiva 2).

con la que mantiene una relación de semejanza interpretativa (Sperber y Wilson, 2000, pp. 274-283; Wilson, 2000; Sperber, 2000; Gutt, 2010; Portolés, 2002, p. 163)[8]; y es también, como cualquier estímulo lingüístico ostensivo, el punto de partida para un proceso –necesariamente humano– de comprensión regulado por el *principio de la pertinencia* (Sperber y Wilson, 1986, p. 198; Ervas, 2010, §2.2.).

En cuanto acto comunicativo derivado de otro previo al que debe *fidelidad interpretativa* (Gutt, 2010, pp. 105-107), toda traducción, ya sea humana o automática, debe garantizar la transmisión del significado explícito e implícito intencionalmente comunicado por el texto origen (principio de metarrepresentación) y debe hacerlo –para que pueda efectivamente hablarse de eficacia y de calidad– garantizando la máxima pertinencia comunicativa en el texto meta (principio de relevancia óptima). El texto traducido ha de aparecer de tal forma a los ojos del lector que la primera interpretación que active en contexto coincida sin esfuerzo añadido con las comunicadas intencionalmente por el texto origen (Gutt, 2010, pp. 40-41). Esto va a suponer necesariamente un proceso de adaptación intralingüística, interlingüística e intercultural (Nord, 1994, pp. 59-67).

Metarrepresentación y *pertinencia óptima* son, pues, condiciones necesarias para que el texto traducido pueda cumplir su función comunicativa en el nuevo entorno o contexto cultural que lo va a acoger. El principio de metarrepresentación está orientado hacia el texto origen y la lengua de partida, y el principio de pertinencia óptima, hacia el texto meta y la lengua de llegada. A la luz del primero, cabe considerar como errores de traducción todos aquellos fenómenos lingüísticos o extralingüísticos que falseen o alteren el significado intencionalmente comunicado

[8] Desde la teoría de la pertinencia se considera que los pensamientos y los enunciados representan un estado de cosas, es decir, una proposición, ya sea real o imaginaria, sobre la que se proyecta una actitud. Los pensamientos son representaciones mentales; los enunciados son representaciones públicas. Pues bien, la noción de *metarrepresentación* surge para dar cuenta de la capacidad que poseemos los seres humanos de representar mentalmente lo que otros piensan (lectura de la mente), así como de representar con una nueva representación pública, es decir, con un nuevo enunciado, lo que otros han dicho. Es el caso de las citas, del discurso reproducido y también de la traducción, que cabe entender como una particular cita interlingüística (Ervas, 2010, §2.2).

Como explica Wilson (2000, p. 411), «A metarepresentation is a representation of a representation: a higher-order representation with a lower-order representation embedded within it». La relación habitual entre ambas no es de equivalencia literal, sino de *semejanza interpretativa.* En palabras de Wilson (2000, p. 426): «Interpretive resemblance is resemblance in content: that is, sharing of implications. Two representations resemble each other (in a context) to the extent that they share logical and contextual implications. The more implications they have in common, the more they resemble each other».

La metarrepresentación constituye, pues, un uso interpretativo del lenguaje en cuanto que no *describe* directamente un estado de cosas, sino que *interpreta* una representación previa a través de una forma proposicional parecida (para la diferencia entre uso descriptivo e interpretativo, véanse Sperber y Wilson 1986, pp. 279-280; Gutt 2010, p. 36). La traducción, como explica Gutt (2010, pp. 105-107), es un caso particular de uso interpretativo interlingüístico. En palabras de Ervas (2010, §2.1): «In intercultural communication, translation – as a paradigmatic way of communicating among cultures – is nothing but a kind of metarepresentation based on an interpretive use of language where it is represented a new, additional utterance built from another one and resembling it in some respects (*Cfr.* Sperber-Wilson, 1986; Gutt, 1991)».

del texto origen. A la luz del segundo, cabe considerar errores de traducción todos aquellos fenómenos lingüísticos o extralingüísticos que atenten contra la naturalidad del texto y del discurso meta generando, en consecuencia, costes cognitivos añadidos no deseados (extrañeza, distracción...) que limitan la calidad de la traducción. Adaptando el principio de la pertinencia a la traducción[9], se puede decir que la calidad del texto traducido está en función de la fidelidad interpretativa a los significados intencionalmente comunicados y del esfuerzo de procesamiento requerido para recuperarlos. La calidad varía, pues, de manera directa a la fidelidad y de manera inversa al esfuerzo. Ambos criterios nos han guiado en la revisión y formulación de la escala de gravedad del error necesaria a fines evaluativos en el marco de la métrica MQM (*Multidimensional Quality Metrics*)[10].

4. PLANTEAMIENTO DE LA INVESTIGACIÓN

4.1. Corpus textual

La investigación se basa en un corpus textual de ámbito académico sobre temas de lingüística constituido por diez resúmenes en italiano traducidos al español y diez resúmenes en español traducidos al italiano, con un total de 1685 y 1645 palabras respectivamente. La extensión concisa de cada resumen nos ha permitido trabajar con textos completos en lugar de muestras[11].

---

[9] Wilson & Sperber (2004, p. 609) explican el principio de pertinencia en los siguientes términos:

(1) Relevance of an input to an individual
a) Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time.
b) Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time (Wilson y Sperber 2004, p. 252).

[10] La métrica MQM se desarrolla entre los años 2012 y 2014 por el Centro de Investigación Alemán para la Inteligencia Artificial (el DFKI) en el marco del proyecto europeo QT Launchpad (Lommel *et al.*, 2014), con el objetivo de intentar armonizar la evaluación de las traducciones humanas y automáticas. Actualmente, se ha consolidado como norma, tanto en el sector de las industrias de la lengua como en la investigación académica. Para la evolución histórica de la evaluación de la calidad de la traducción (TQE, por sus siglas en inglés), remitimos a Lommel *et al.* (2014) y Moorkens *et al.* (2018). La página oficial está disponible en el siguiente enlace: https://themqm.org/.

[11] El tamaño de la muestra es una de las decisiones más importantes que debe tomar el investigador (Sánchez Ramos y Rico Pérez, 2020). En la práctica de la evaluación humana de la calidad de la traducción se reconocen tres dimensiones de muestreo: pequeña (menos de quinientas palabras), media (entre quinientas y cinco mil palabras) y grande (más de cinco mil palabras) (Lommel *et al.*, 2024). El uso de muestras de tamaño pequeño no se considera fiable, ya que el acuerdo entre anotadores es demasiado bajo y requiere métodos complejos de control estadístico de la calidad; de ahí la necesidad de utilizar muestras de tamaño medio o grande. De hecho, como explican Lommel *et al.* (2024), «Evaluators frequently work with samples in the range of 500 to 20,000 words, depending on the size of the project and the resources available for evaluation». Ahora bien, si se persiguen resultados

Desde el punto de vista de la tipología textual, el resumen es un género discursivo de extensión breve característico del discurso académico. Presenta en clave sintética las ideas principales desarrolladas en el trabajo científico que introduce (artículo de revista, capítulo de libro, etc.). Su función es fundamentalmente informativa. Su modo discursivo es expositivo y posee características propias directamente deudoras de la función y del campo de especialización: orden, claridad, síntesis, cohesión, rigor, precisión léxica, presencia de tecnicismos y ausencia de subjetividad (Mapelli, 2018, pp. 101-121).

4.2. Metodología

La investigación compara el rendimiento de cuatro sistemas de traducción automática: DeepL y Google Translate (GoogleT) para TAN, y ChatGPT-4o y Deepseek, para LLM. La metodología se basa en el análisis de errores, según la métrica MQM, que tiene dos pilares fundamentales: la tipología y la gravedad del error[12]. La unidad de análisis es la frase, entendida como el segmento lingüístico comprendido entre dos puntos, si bien hemos tenido en cuenta la totalidad del texto para evaluar la gravedad de los errores hallados. La anotación, completamente humana, ha sido realizada con el programa informático UAMCorpusTool (O'Donnell, 2008, 2021) y sometida a dos procesos de revisión por pares[13]. Para generar las traducciones con los LLM, hemos utilizado un *prompt* básico (*Traduce al español*; *Traduci in italiano*), entendido como punto de partida para futuras investigaciones sobre el impacto del diseño de *prompts*. Las traducciones las generamos el día 18 de abril del 2025.

4.3. Criterios de evaluación (cuantitativa y cualitativa)

4.3.1. *Taxonomía de error*

La métrica MQM distingue siete posibles categorías de error con sus correspondientes subcategorías[14]. Indicamos a continuación las categorías que se han revelado útiles para nuestro estudio por la propia naturaleza del tipo de texto (cuadro 1).

---

con poder de generalización estadística, diversos estudios recientes (por ejemplo, Gladkoff *et al.*, 2022) han demostrado que incluso muestras de dos mil palabras introducen un índice de incertidumbre significativo, motivo por el cual proponen un tamaño mínimo de cuatro mil palabras.

[12] Para la asignación de la categoría de error, hemos seguido el árbol de decisión (*decision tree*) que figura en la página *web* oficial de MQM: https://themqm.org/error-types-2/decisiontree/.

[13] Aunque se han desarrollado métricas automáticas para la evaluación de calidad, «... human translation quality evaluation is more than ever THE Golden Standard of measurement and benchmarking for quality measurement, since it is the only reliable way to validate any automatic translation quality evaluation» (Lommel *et. al.*, 2024).

[14] Para la adecuada comprensión de las categorías, tal y como han sido concebidas por los grupos de investigación implicados en la creación de la métrica MQM, remitimos a las definiciones y

CUADRO 1. CATEGORÍAS DE LA MÉTRICA MQM EMPLEADAS EN LA ANOTACIÓN DE LOS ERRORES

Categoría	Descripción	Subcategoría
Terminología (*Terminology*)	El término del texto meta (TM) no se ajusta a las expectativas terminológicas del campo de especialización o no es el equivalente normativo correcto del término utilizado en el texto origen (TO).	Variación terminológica no justificada Tecnicismo incorrecto
Precisión (*Accuracy*)	El TM no se corresponde con exactitud con el contenido proposicional del TO.	Traducción errónea > *Falso sentido*[1] > TM ambiguo > Falso amigo > Demasiado literal > Alucinación Sobretraducción Subtraducción Adición Omisión No traducir Sin traducir Error factual arrastrado (desde el TO) *Error material arrastrado*[2] (desde el TO)
Convenciones lingüísticas - fluidez (*Linguistic conventions*)	La traducción no respeta las normas gramaticales y de buena formación del texto en la lengua meta.	Gramática > *Gramática* > Colocación Tipografía Ortografía Convenciones textuales > Coherencia > Cohesión
Estilo (*Style*)	La formulación es gramaticalmente correcta, pero no se ajusta a las convenciones estilísticas y de registro propias de la tipología textual y de la tradición discursiva del texto en la lengua de llegada.	Registro No natural
Entorno cultural (*Audience appropriateness*)	La traducción no resulta funcional porque no se ajusta al entorno cultural del TM o porque podría resultar ofensiva.	(Traducción) no inclusiva
Diseño y composición de página (*Design and markup*)	La traducción no respeta algún aspecto relacionado con la diagramación de la página. Son errores relacionados con la tipografía y formato de caracteres, la disposición de párrafos, la inserción de elementos gráficos, etcétera).	Formato de caracteres

[1] Por motivos de claridad descriptiva, precisamos la voz *Traducción errónea (Mistranslation)* con la subcategoría *Falso sentido* para dar cuenta de un buen número de errores que no se adaptan a ninguna de las subcategorías ya reconocidas por la métrica. Entendemos por *Falso sentido* el error que «... resulta de una mala interpretación del sentido de una palabra o de un enunciado en un contexto dado, sin llegar a causar contrasentido o sin sentido» (Hurtado Albir, 2001, p. 291, que sigue, a su vez, a Delisle, 1993, p. 31). Hacemos lo mismo con la subcategoría *Gramática* en *Convenciones lingüísticas*.

[2] Proponemos la subcategoría *Error material arrastrado* para dar cuenta de los errores de traducción provocados por la presencia de un error material (es decir, una errata) en el texto origen. Se diferencia así del error factual, que es el provocado por la referencia a hechos claramente falsos en el texto origen.

---

ejemplos propuestos en la página web oficial del modelo: https://themqm.org/the-mqm-full-typology/. Hasta donde nuestro conocimiento alcanza, no existe una traducción oficial consensuada al español o al italiano, algo que sería muy deseable. Traducimos *Audience appropriateness* con *Entorno cultural*, término que tomamos de Duro Moreno (2012).

4.3.2. *Gravedad del error*

El nivel de gravedad refleja el «... efecto de un error específico en la usabilidad del texto [...] según el grado de riesgo que suponga para la calidad de la traducción» (MQM Council, 2025; la traducción es nuestra). La escala propuesta por la métrica MQM distingue cuatro niveles (neutro, menor, mayor, crítico).

Para la presente investigación, hemos revisado la escala desde la perspectiva de la teoría de la pertinencia. Así, reconocemos solo tres niveles de error: menor, mayor, crítico. El cuarto nivel lo sustituimos con el de *variante estilística*, en cuanto que no lo consideramos constitutivo de error. El modelo revisado se adapta mejor al campo y la tipología textual de nuestro corpus de análisis y a los fenómenos hallados y que buscamos describir. Los criterios aplicados son los siguientes:

- falseamiento de la explicatura o de la implicatura intencionalmente comunicadas: nos preguntamos si la traducción falsea o no el significado explícita o implícitamente comunicado;
- adición de costes de procesamiento no presentes en el original: nos preguntamos si la traducción añade o no costes de procesamiento que restan naturalidad y relevancia al texto;
- identificación del error: nos preguntamos si el error es o no identificable sin consultar el texto origen.

La escala (cuadro 2) queda como sigue (para mayor claridad, añadimos un ejemplo):

CUADRO 2. ESCALA DE GRAVEDAD DEL ERROR REVISADA

Nivel de gravedad	Definición
Error menor	No falsea el significado intencionalmente comunicado por el TO, pero añade costes mínimos de procesamiento en el texto meta (TM) (como errores de deletreo, acentuación, puntuación, traducción indebida de títulos, etc.). No afecta a la fiabilidad del contenido y muy limitadamente a la comprensibilidad del texto. Ej. El sistema traduce *ordenadores discursivos* con *ordinatori *discursivi* en lugar de *discorsivi.*
Error mayor	Falsea el significado intencionalmente comunicado por el TO o viola las normas de buena formación del texto meta (TM). Provoca extrañeza y añade costes de procesamiento elevados. Afecta a la fiabilidad o a la comprensibilidad del texto, pero es identificable. Ej. GoogleT traduce *adverbio de lugar* como *avverbio locale* en lugar de *avverbio di luogo*.
Error crítico[3]	Falsea el significado intencionalmente comunicado, pero no genera costes de procesamiento porque resulta plausible y pasa desapercibido. Afecta a la fiabilidad del contenido. Solo puede ser identificado consultando el TO. Ej. DeepL no resuelve adecuadamente la relación anafórica señalada por el determinante demostrativo *esta* y falsea la explicatura. TO: El estudio presenta los datos de 20 hablantes nativos de italiano que visualizaron una entrevista en español subtitulada *en esta misma lengua*. DeepL: Lo studio presenta i dati di 20 parlanti nativi italiani che hanno visto un'intervista in spagnolo sottotitolata *in italiano*.

[3] Según la métrica MQM, el error *crítico* es aquel que vuelve el texto inutilizable para el propósito con el que fue concebido o que supone un riesgo grave de daño físico, financiero o reputacional. En nuestro caso, por la propia naturaleza de los textos analizados, todos los errores mayores y críticos afectarían a la reputación no solo del autor, sino también del traductor y de la revista o volumen que los acogiera; de ahí que hayamos enfocado el error crítico desde una perspectiva cognitiva.

El concepto de *variante estilística* que proponemos en sustitución del de error *neutro* de la métrica MQM no es un mero cambio de denominación, sino que responde a una decisión metodológica con la que pretendemos dar cuenta del alto índice de variación sin error observado en la comparación entre los sistemas automáticos, los cuales se aproximan cada vez más al traductor humano: «… translation as interlingual interpretive use allows for too much variation» (Gutt 2010, 130). En el caso concreto del texto académico, la variación estilística (cuadro 3) se confirma, además, como una dimensión particularmente relevante en cuanto analizable a la luz de la máxima o imperativo de claridad y de las disposiciones del *lenguaje claro*, como veremos en §5.

Cuadro 3. Definición de variante estilística	
Variante estilística	Variante de formulación compatible con las expectativas de estilo y de registro previstas por la tipología textual y la tradición discursiva. No es un error. Ej. Los cuatro sistemas ofrecen variantes adecuadas para la traducción de la partícula que introduce el tema (*a proposito di*) y de la subordinada de participio *da noi adottada*. TO: *A proposito della* suddivisione *da noi adottata*. GoogleT: *Respecto a* la subdivisión *que hemos adoptado.* DeepL: *Con respecto a* la subdivisión *que adoptamos*. ChatGPT: *En relación con* la clasificación *adoptada*. Deepseek: *Respecto a la* clasificación *adoptada*.

4.3.3. *Puntuación de calidad*

La métrica MQM dispone de distintas fórmulas para calcular una puntuación numérica representativa de la calidad de la traducción (Lommel *et al.*, 2024). En nuestro estudio aplicamos una fórmula lineal sin calibración que pasa por tres fases.

Primero, se suma el número de errores multiplicado por el peso numérico asignado a cada *nivel de gravedad*, según la siguiente escala (*Absolute Penalty Total*):

- Errores menores: 1

- Errores mayores: 3

- Errores críticos: 9[15].

Segundo, se divide este valor por el *número de palabras totales* generado por cada sistema (*Per-Word Penalty Total*) (cuadro 4).

[15] En línea con las directrices del modelo MQM (Lommel *et al.* 2024, 15), reasignamos los pesos para adecuar la escala al tipo de texto, manteniendo una relación exponencial que, en nuestro caso, es de base 3.

CUADRO 4. NÚMERO DE PALABRAS TOTALES GENERADO POR CADA SISTEMA	
ESPAÑOL > ITALIANO	ITALIANO > ESPAÑOL
GoogleT: 1528	GoogleT: 1750
DeepL: 1551	DeepL: 1790
ChatGPT: 1615	ChatGPT: 1817
Deepseek: 1528	Deepseek: 1757

Tercero, se traslada el resultado a una escala de 0 a 100 para obtener la puntuación de calidad MQM (*Raw Quality Score*) (cuadro 5).

Step	Raw Quality Score Calculation	Formulas
1	Absolute Penalty Total (APT)	$\sum_{i,j} Error\ Count_{ij} \times Severity\ Multiplier_j \times Error\ Type\ Weight_i$ Where: *i* = index for Error Types, *j* = index for Severity Level.
2	Per-Word Penalty Total (PWPT)	$\frac{Absolute\ Penalty\ Total}{Evaluation\ Word\ Count}$
3	Raw Quality Score (RQS)	$1 - PWPT$ (for a scale with a maximum of 1) or $100 - (PWPT \times 100)$ (for a scale with a maximum of 100)

Cuadro 5. Fuente: MQM Council.

Presentamos a continuación los resultados tanto desde un punto de vista cuantitativo como cualitativo.

5. RESULTADOS

5.1. ANÁLISIS CUANTITATIVO COMPARADO

El análisis cuantitativo constituye una base empírica y numérica que, por un lado, respalda el análisis cualitativo y, por otro lado, proporciona una visión de conjunto que permite comparar el impacto de las distintas categorías de error.

5.1.1. *Número y tipología de errores*

Las tablas 1 y 2 recogen el número total de errores por sistema y el porcentaje de cada categoría de error.

TABLA 1. TOTAL DE ERRORES ABSOLUTOS POR SISTEMA (ES>IT)								
Español > Italiano								
Categoría MQM	ChatGPT		Deepseek		DeepL		GoogleT	
	n.º	%	n.º	%	n.º	%	n.º	%
Terminología	8	47,06	6	23,08	20	41,67	23	29,49
Precisión	9	52,94	16	61,54	23	47,92	43	55,13
Fluidez	-	0,00	3	11,54	0	0,00	9	11,54
Estilo	-	0,00	-	0,00	4	8,33	2	2,56
Entorno cultural	-	0,00	-	0,00	1	2,08	1	1,28
Diseño y composición	-	0,00	1	3,85	-	0,00	-	0,00
Total	17	100,00	26	100,00	48	100,00	78	100.0

TABLA 2. TOTAL DE ERRORES ABSOLUTOS POR SISTEMA (IT>ES)								
Italiano>Español								
Categoría MQM	ChatGPT		Deepseek		DeepL		GoogleT	
	n.º	%	n.º	%	n.º	%	n.º	%
Terminología	4	33,33	7	23,33	14	33,33	12	24,49
Precisión	8	66,67	14	46,67	17	40,48	21	42,86
Fluidez	-	0,00	8	26,67	10	23,81	13	26,53
Estilo	-	0,00	-	0,00	1	2,38	3	6,12
Entorno cultural	-	0,00	-	0,00	-	0,00	-	0,00
Diseño y composición	-	0,00	1	3,33	-	0,00	-	0,00
Total	12	100,00	30	100,00	42	100,00	49	100,00

La categoría *Precisión* es la que presenta el mayor número de errores en ambas direcciones y en todos los sistemas, seguida de *Terminología* y *Fluidez*. ChatGTP es el que resulta más eficaz, y GoogleT, el que muestra un peor rendimiento. En la dirección español-italiano, resulta más clara la diferencia entre los LLM y los TAN, a favor de los primeros.

Ilustramos, a continuación, el análisis de cada categoría de error por separado. La tabla 3 refleja el total de errores en ambas direcciones en la categoría *Terminología*. En ambas direcciones, los LLM resultan más eficaces que los TAN en la identificación y traducción de tecnicismos. La diferencia resulta más marcada en la dirección español>italiano.

TABLA 3. ERRORES DE LA CATEGORÍA *TERMINOLOGÍA* POR SISTEMA EN LAS DOS DIRECCIONES								
	Español>Italiano				Italiano>Español			
Terminología	ChatGPT	Dpseek	DeepL	GoogleT	ChatGPT	Dpseek	DeepL	GoogleT
Variación terminológica	-	-	-	-	-	-	2	-
Tecnicismo incorrecto	8	6	20	23	4	7	12	12
Total	8	6	20	23	4	7	14	12

En la tabla 4 figuran los errores de la categoría *Precisión,* que surgen cuando la traducción no garantiza una correspondencia precisa con el significado proposicional o explicatura del texto origen. Como sucedía para la categoría *Terminología*, los LLM se muestran de nuevo más eficaces que los TAN en ambas direcciones. GoogleT es el sistema que muestra un peor rendimiento, sobre todo en italiano. En la dirección italiano-español, los errores se concentran en la categoría *Traducción errónea*. En la dirección español-italiano, se añade, además, *No traducir* y destaca GoogleT en errores de *Falso sentido* y de *Alucinación*.

TABLA 4. ERRORES DE LA CATEGORÍA *PRECISIÓN* POR SISTEMA EN LAS DOS DIRECCIONES								
	Español>Italiano				Italiano>Español			
Precisión	ChatGPT	Dpseek	DeepL	GoogleT	ChatGPT	Dpseek	DeepL	GoogleT
Traducción errónea	1	2	7	26	2	8	9	14
> Falso sentido	-	-	4	14	2	6	7	12
> TM ambiguo	1	1	1	3	-	1	1	-
> Falso amigo	-	-		1	-	-	-	-
> Demasiado literal	-	-	-	-	-	-	1	-
> Alucinación	-	1	2	8	-	1	-	2
Sobretraducción	-	-	-	-	-	-	-	-
Subtraducción	3	2	2	3	1	2	2	2
Adición	-	-	-	2	-	1	-	-
Omisión	-	-	1	1	1	1	1	1
No traducir	4	12	12	10	3	2	3	3
Sin traducir	1	-	1	-	1	-	-	-
Error factual arrastrado	-	-	-	-	-	-	1	-
Error material arrastrado	-	-	-	1	-	-	1	1
TOTAL:	9	16	23	43	8	14	17	21

En cuanto a la categoría *Fluidez* (tabla 5), los errores son en su mayoría *menores* (por motivos gramaticales, ortográficos y tipográficos). En ambas direcciones, ChatGTP se sitúa como el sistema más eficaz, y GoogleT, como el de rendimiento menos satisfactorio.

TABLA 5. ERRORES DE *CONVENCIONES LINGÜÍSTICAS* (FLUIDEZ) POR SISTEMA EN LAS DOS DIRECCIONES								
	Español>Italiano				Italiano>Español			
Convenciones lingüísticas (fluidez)	ChatGPT	Deepseek	DeepL	GoogleT	ChatGPT	Deepseek	DeepL	GoogleT
Gramática	-	-	-	-	-	1	1	2
> Gramática	-	-	-	-	-	-	-	1
> Colocación	-	-	-	-	-	1	1	1
Tipografía	-	-	-	-	-	6	7	3
Ortografía	-	3	-	3	-	-	1	3
Convenciones textuales	-	-	-	-	-	1	1	5
> Coherencia	-	-	-	1	-	-	-	1
> Cohesión	-	-	-	5	-	1	1	4
TOTAL	0	3	0	9	0	8	10	13

En §4 hemos definido la *variación estilística* como una variante de formulación aceptable en cuanto compatible con las expectativas de estilo y registro previstas por la tipología textual y la tradición discursiva. Cabe hablar, sin embargo, de error cuando la traducción no respeta dichas expectativas y provoca, en consecuencia, costes de procesamiento que interrumpen la lectura y la comprensión del texto. Transgredir las convenciones estilísticas y de registro fijadas por la tradición y la costumbre repercute directamente en la eficacia comunicativa (Carrillo Guerrero, 2005). Los LLM no producen errores en ninguna dirección.

TABLA 6. DETALLES DE ERRORES DE LA CATEGORÍA *ESTILO* POR SISTEMA EN LAS DOS DIRECCIONES								
	Español>Italiano				Italiano>Español			
Estilo	ChatGPT	Deepseek	DeepL	GoogleT	ChatGPT	DeepL	Deepseek	GoogleT
Registro	-	-	2	2	-	-	1	3
Poco natural	-	-	2	-	-	-	-	-
TOTAL:	0	0	4	2	0	0	1	3

Bajo la subcategoría *Entorno cultural* (tabla 7) se ha documentado un único error en DeepL y GoogleT, cualitativamente pertinente porque pone de manifiesto el reto que supone la traducción del discurso inclusivo para los sistemas neuronales. Lo comentaremos en la próxima sección. Por lo que se refiere a los errores de *Diseño*, hemos documentado solo un error de Deepseek en ambas direcciones, referido al uso inapropiado de la negrita.

TABLA 7. DETALLES DE ERRORES DE LA CATEGORÍA *ENTORNO CULTURAL* POR SISTEMA EN LAS DOS DIRECCIONES								
	Español>Italiano				Italiano>Español			
Entorno cultural	ChatGPT	Deepseek	DeepL	GoogleT	ChatGPT	DeepL	Deepseek	GoogleT
No inclusivo	-	-	1	1	-	-	-	-
TOTAL:	-	-	1	1	-	-	-	-

5.1.2. *Gravedad del error*

La tabla 8 recoge el total de errores por cada sistema, distribuido por nivel de gravedad.

TABLA 8. TOTAL DE ERRORES DISTRIBUIDOS POR NIVEL DE GRAVEDAD								
	Español>Italiano				Italiano>Español			
	ChatGPT	Deepseek	DeepL	GoogleT	ChatGPT	Deepseek	DeepL	GoogleT
Crítico	6	12	18	25	5	8	7	11
Mayor	7	6	18	37	3	9	19	23
Menor	4	8	12	16	4	13	16	15
TOTAL:	17	26	48	78	12	30	42	49

En ambas direcciones, se observa una diferencia significativa en la distribución de errores graves (críticos y mayores) entre los sistemas generativos y no generativos. Hacia el italiano, ChatGTP y Deepseek producen respectivamente 13 y 18 errores graves (mayores y críticos), frente a los 36 y 62 errores producidos por DeepL y GoogleT. Hacia el español, el número total de errores desciende, pero se advierte igualmente el mejor rendimiento de los LLM: ChatGTP y Deepseek generan respectivamente 8 y 17 errores graves, frente a los 26 y 34 errores de DeepL y GoogleT. En ambas direcciones, ChatGPT destaca por ser el sistema que no solo produce menos errores totales, sino también una menor cantidad de errores graves (críticos y mayores). Por su parte, GoogleT muestra una tendencia opuesta, caracterizada por un número comparativamente elevado de errores y por una mayor concentración de errores críticos y mayores, lo que incide de manera negativa en su rendimiento global.

5.1.3. *Puntuación MQM*

La tabla 9 recoge la puntuación de calidad MQM obtenida por los distintos sistemas. La puntuación refleja la porción de texto libre de errores.

TABLA 9. PUNTUACIÓN MQM			
Español>Italiano		Italiano>Español	
Sistema	Puntuación MQM	Sistema	Puntuación MQM
ChatGPT	95,11	ChatGPT	96,81
Deepseek	91,33	Deepseek	93,63
DeepL	85,30	DeepL	92,40
GoogleT	76,96	GoogleT	89,54

La puntuación MQM confirma los resultados obtenidos en los apartados anteriores y sitúa a los sistemas generativos por delante de los no generativos en la traducción automática del texto académico en ambas direcciones. Cabe observar lo siguiente:

1. Los LLM muestran un rendimiento superior al de los TAN;
2. Los LLM generan menos errores graves (críticos y mayores) que los TAN;
3. Dentro de los generativos, ChatGPT 4-o muestra un rendimiento claramente superior al de Deepseek;
4. GoogleT es el sistema que produce peores resultados;
5. En todos los sistemas y en ambas direcciones, los errores más frecuentes resultan ser los de *Precisión* y *Terminología*, categorías de errores relacionadas con la recuperación de la explicatura de nivel inferior y esperables en un texto y discurso especializado como es el académico;
6. Los errores de *Alucinación* se documentan tanto con los LLM como con los TAN, siendo GoogleT, en la traducción hacia el italiano, el que más fenómenos de esta categoría ha generado;
7. Todos los sistemas ofrecen mejores resultados hacia el español.

5.2. Análisis cualitativo comparado

Como decíamos en §3, los errores de metarrepresentación afectan la fiabilidad del contenido traducido. En la métrica MQM se corresponden principalmente con las categorías *Terminología, Precisión* y *Entorno cultural.* Más directamente relacionados con el principio de pertinencia óptima y la optimización del estímulo ostensivo son los errores que afectan a las categorías *Fluidez*, *Estilo* y *Diseño y composición*. Estos últimos están relacionados con la comprensibilidad del texto y se interpretan en términos de costes de procesamiento añadidos.

Abordamos a continuación el análisis cualitativo de los errores automáticos encontrados. Nos interesa describir y explicar los distintos tipos de error documentados, así como identificar fenómenos y patrones específicos de traducción automática que no se verifican en la humana. Intentamos, de este modo, adentrarnos en los mecanismos de razonamiento y pensamiento (no humano) de los modelos avanzados[16].

5.2.1. *Errores relacionados con la metarrepresentación de la explicatura*

Como señalábamos en §3 y §4.1., por la propia naturaleza del tipo de texto, los errores más numerosos y graves son los que remiten a lo que el ser humano entiende como *metarrepresentación fallida de la proposición o explicatura de nivel inferior*, bien porque se propone un tecnicismo inadecuado o no equivalente (error de *Terminología*), bien porque el texto no refleja con exactitud el significado proposicional original (error de *Precisión*). Pese a su origen puramente algorítmico, la mente humana los interpreta como intentos no logrados de descodificación, desambiguación, saturación referencial o enriquecimiento pragmático. En algunos casos (menos), el error afecta a la metarrepresentación de la modalidad oracional y la fuerza ilocutiva. Señalamos a continuación los fenómenos de mayor interés en relación con las distintas categorías de error.

5.2.1.1. Terminología

La terminología es una dimensión fundamental de los discursos especializados y, por ende, de la traducción especializada, dado que refleja los conocimientos de un determinado campo (Montero Martínez *et al.*, 2011, p. 24). Como señala Calvi (Calvi *et al.*, 2011, pp. 27-28): «reconocerla e interpretarla correctamente es el primer paso para trasladar un mensaje de una lengua a otra». Los errores de terminología pueden afectar tanto a lexemas, colocaciones o construcciones técnicas monosémicas como a sentidos técnicos desarrollados por unidades léxicas de uso común. Así, por ejemplo, *morfema* es un tecnicismo monosémico de la gramática; *adjunto* y *aspecto* son en cambio dos palabras de uso común (documento *adjunto*, *aspecto* familiar) que asumen un sentido especializado en el campo de la sintaxis; *adverbio temporal* es una colocación frecuente (equivalente a *adverbio de tiempo*), pero no podemos decir lo mismo de *adverbio local*, pese a ser un sintagma formalmente idéntico. Como decíamos en la sección anterior, los errores de *Terminología* son más numerosos en los TAN y se han documentado con más frecuencia hacia el italiano.

Así, por ejemplo, *adjuntos* se traduce como *aggettivi* en lugar de *aggiunti* o *complementi* (DeepL); *pronombres personales átonos,* como *pronomi personali non sottolineati* (DeepL) en lugar de *atoni*; *formas desdobladas*, como *forme divise* (GoogleT) en

[16] Los conceptos de *pensamiento* y *razonamiento* han de entenderse en sentido figurado en cuanto referidos a procesos y mecanismos puramente algorítmicos. Lo mismo cabe decir de *inteligencia* artificial, *conciencia* artificial o *capacidad inferencial*, habituales en la bibliografía sobre el tema.

lugar de *forme sdoppiate o scisse*; *realce prosódico*, como *esaltazione prosodica* (GoogleT) en lugar de *rilievo prosodico.*

Los errores de *Terminología* afectan también a los acrónimos, que acceden a la lengua meta a través del filtro del inglés. Es el caso de SFL (*Spanish Foreign Language*) (GoogleT) en lugar de ELE. Y a construcciones de significado técnico que el sistema no reconoce. Por ejemplo, *marcas de evidencialidad* (TO) pasa como *indicatori di prova* (GoogleT) o como *segni di evidenza* (DeepL) en lugar de *marcatori di evidenzialità* (ChatGTP y Deepseek).

Los errores de *Terminología* se documentan también en la dirección italiano-español. Así, por ejemplo, el sustantivo *indefinitezza* pasa erróneamente como *indefinición* (GoogleT, DeepL, ChatGTP) o *indefinitez* (Deepseek) en lugar de *indefinitud*, que es el tecnicismo utilizado en el campo de la lingüística para expresar 'condición de indefinido' con referencia al artículo. Particularmente interesante por tratarse de un sentido *ad hoc* es el caso del adjetivo italiano *pieno* en *sintagmi nominali pieni.* Los sistemas neuronales no reconocen el sentido especializado ('significado conceptual' o 'léxico' frente a 'gramatical') y traducen como *sintagmas nominales completos,* falseando la explicatura. Véase el ejemplo (1).

	EJEMPLO (1)
TO	può essere realizzato non solo da *sintagmi nominali pieni* [...] ma anche da espressioni lessicalmente vuote.
Google y Deepl	puede lograrse no sólo mediante sintagmas nominales **completos* [...] sino también mediante expresiones léxicamente vacías.
ChatGPT	puede realizarse no solo mediante sintagmas nominales *plenos* [...], sino también mediante expresiones léxicamente vacías, como los pronombres.
Deepseek	puede realizarse no solo mediante sintagmas nominales *léxicamente llenos* [...] sino también a través de expresiones léxicamente vacías, como los pronombres.

Una buena parte de los errores terminológicos documentados en la dirección italiano-español afectan a los tecnicismos desarrollados para dar cuenta de las cuestiones relacionadas con el *lenguaje inclusivo*. Así, encontramos *masculino sobredimensionado* (GoogleT), *sobreextendido* (DeepL y Deepseek) y *sobregeneralizado* (ChatGPT) en lugar de *masculino genérico*; *femenino exagerado* (GoogleT) en lugar de *femenino genérico*; *masculinidad inclusiva* (GoogleT) en lugar de *masculino inclusivo*.

Asimismo, el análisis cualitativo ha permitido advertir tres tendencias de traducción automática que conducen a error y que responden claramente a mecanismos de razonamiento no humano:

1. *Traducción por semejanza formal*: el sistema propone un término formalmente parecido, aunque semánticamente no pertinente. La tendencia se observa solo en los sistemas de TAN. Así, por ejemplo, DeepL propone *(valori) formulaci* por *formulativi* (TO: *valores formulativos*); *l'elemento fonico* por *forico* (TO: elemento *fórico)*, GoogleT propone *complemento proposizionale* en lugar de *complemento di proporzione* (TO: complemento de *proporción*).

2. *Alejamiento innecesario de la literalidad*: son aquellos casos en los que la traducción literal del tecnicismo original hubiera sido la opción más adecuada y también la más sencilla e inmediata para el traductor humano. Es el caso de *encargos de traducción* (TO), que Deepseek traduce como *commissioni traduttive* en lugar de *incarichi di traduzione;* o el sintagma *efectos afectivos no proposicionales,* que DeepL traduce como *effetti affettivi non propositivi* en lugar de *effetti affettivi non proposizionali*; o el error arriba mencionado de *pronomi personali non sottolineati* (DeepL) en lugar de *atoni* (TO: *átonos*).
3. *Transferencia de tecnicismos de otros campos de especialización*: se observan casos en los que los sistemas tienden a emplear tecnicismos existentes, pero pertenecientes a otros ámbitos de especialidad. Así, el *neomorfema* -e (TO) se ha traducido al italiano con *neomorfismo* -e (DeepL), un tecnicismo perteneciente al ámbito del diseño; *marcación (discursiva)* (TO) se ha traducido con *etichettatura* (DeepL), un tecnicismo del ámbito comercial (producción de productos a gran escala).

5.2.1.2. Precisión

Los errores de *Precisión* son más numerosos en los TAN en ambas direcciones. La categoría *Precisión* incluye todas aquellas traducciones que no reflejan con exactitud la explicatura original, bien porque generan una ambigüedad que no existía en el texto origen, bien porque falsean de alguna manera el significado original. Así, en (2), la traducción literal del adjetivo *normal* propuesta por los TAN genera en italiano una ambigüedad inexistente en el TO y no fácil de resolver en contexto: *normale* entendido como 'normativo' y *normale* entendido como 'habitual'. Los LLM, en cambio, proponen el lexema *corrente*, el cual no da lugar a ambigüedad:

	EJEMPLO (2)
TO	En su empleo *normal* en la interacción espontánea actual, la unidad ha desarrollado valores conclusivos, formulativos y modales
GoogleT	Nel suo *normale* utilizzo nell'interazione spontanea attuale
DeepL	Nel suo uso *normale* nell'interazione spontanea corrente
ChatGPT	Nell'uso *corrente* all'interno dell'interazione spontanea
Deepseek	Nel suo impiego *corrente* nell'interazione spontanea

Los ejemplos (3) y (4) son casos de falseamiento de la explicatura. En (4), la adición de *entrambe* (GoogleT) da a entender que las ediciones del *Diccionario de uso* de María Moliner son solo dos. En (5), *fornisce* (GoogleT) en lugar de *inverte*:

	EJEMPLO (3)
TO	se ha decidido compilar un corpus [...] para luego contrastarlo tanto con las *posteriores ediciones* del mismo DUE
GoogleT	si è deciso di compilare un corpus [...] per confrontarlo con *entrambe le edizioni successive* dello stesso DUE
DeepL	si è deciso di compilare un corpus [...] e di confrontarlo poi sia con le edizioni successive del DUE

ChatGPT	si è deciso di compilare un corpus [...] per poi confrontarlo sia con le edizioni successive dello stesso DUE
Deepseek	si è deciso di compilare un corpus [...] per poi confrontarlo sia con le successive edizioni dello stesso DUE

EJEMPLO (4)	
TO	Este operador *invierte* la orientación discursiva de la argumentación.
GoogleT	Questo operatore *fornisce* la direzione discorsiva dell'argomentazione.

Véase, a continuación, el ejemplo (5). Los LLM representan fielmente el significado original. En cambio, GoogleT falsea la explicatura y DeepL genera, además, una ambigüedad inexistente en el texto origen debido a la posibilidad de interpretar *medio* como adjetivo o como adverbio y de segmentar, en consecuencia, de dos maneras: [[ciudadano medio] culto] o [ciudadano [medio culto]]:

EJEMPLO (5)	
TO	[...] un tipo di testo che [...] ha il dovere di essere chiaro e comprensibile *anche* per il *cittadino di media istruzione*.
GoogleT	[...] un tipo de texto que [...] tiene el deber de ser claro y comprensible *incluso* para el *ciudadano medio*.
DeepL	[...] un tipo de texto que [...] tiene el deber de ser claro y comprensible *incluso* para el *ciudadano medio culto*..
ChatGPT	[...] un tipo de texto que [...] tiene el deber de ser claro y comprensible para un *ciudadano con un nivel medio de instrucción*.
Deepseek	...] un tipo de texto que [...] debe garantizar claridad y comprensibilidad *incluso* para *ciudadanos con un nivel educativo medio*.

Obsérvese, por otro lado, que los errores pueden afectar también a las partículas discursivas que guían la labor inferencial del lector, tanto a las que funcionan desde la periferia oracional, (los marcadores discursivos propiamente dichos) como a las que están sintácticamente integradas en los constituyentes oracionales y orientan la obtención de significados implícitos que contribuyen a la explicatura. Es el caso, por ejemplo, de los adverbios de foco como *también* o *hasta* en español o como *anche* y *perfino* en italiano (Portolés, 2002; 2016, p. 692). En el ejemplo anterior (5), tenemos precisamente un error en la traducción de *anche*. El adverbio destaca el segmento *per il cittadino di media istruzione* (foco). De este modo, corrige el supuesto de que la guía tuviera el deber de ser clara y comprensible solo para los redactores y coloca en el mismo nivel a ambas categorías de lectores: tanto los redactores como los ciudadanos con un nivel medio de instrucción. Su equivalente contextual más próximo es el adverbio español *también*. GoogleT, DeepL y Deepseek traducen, sin embargo, como *incluso,* imponiendo así, convencionalmente, una implicatura escalar inexistente en el original. ChatGTP prescinde de la marcación.

Terminamos con los dos últimos ejemplos. La omisión de una coma es a veces irrelevante; en otras ocasiones, puede distorsionar el significado, transformando, como en (6), el sintagma *declinati al femminile* en un complemento especificativo:

EJEMPLO (6)	
TO	i cosiddetti nomina agentis, cioè *i nomi professionali, declinati al femminile* (ministra, assessora, ecc.).
ChatGPT	los llamados nomina agentis, es decir, *los nombres de profesión en su forma femenina* (ministra, asesora, etc.).
Deepseek	los llamados nomina agentis, es decir, *los nombres profesionales en su forma femenina* (ministra, assessora, etc.).

En (7), la omisión del signo de interrogación falsea la fuerza ilocutiva del acto de habla original, que pasa de pregunta a aseveración:

EJEMPLO (7)	
TO	i) ¿Existe un uso innovador de las comillas en los titulares de prensa española? ii) ¿Existe una relación entre este uso y las prácticas actuales de lectura y escritura?
DeepL	i) esiste un uso innovativo delle virgolette nei titoli dei giornali spagnoli; ii) esiste una relazione tra questo uso e le attuali pratiche di lettura e scrittura

Como para la terminología, se han advertido cuatro tendencias específicamente automáticas reconducibles a un razonamiento no humano de base algorítmica:

1. *Traducción errónea por reanálisis composicional*: el sistema altera arbitrariamente las relaciones entre los constituyentes y descodifica, en consecuencia, de manera incorrecta. Cuando es necesario, cambia la categoría y las marcas de flexión de los elementos léxicos. Así, un sintagma como *la investigación sobre marcadores discursivos y variación* (TO) se convierte en *la ricerca sui marcatori e la variazione del discorso*, donde el adjetivo *discursivos* se reinventa como complemento preposicional de *variación* (GoogleT); el sintagma *lingua standard-dialetto* (TO) se convierte en *lengua-dialecto estándar* (DeepL). En el ejemplo (8), tenemos dos sintagmas nominales coordinados: *diverso livello interlinguistico e differenti L1*. En la traducción al español, el sustantivo *livello* se flexiona en plural para concordar con *diferentes* y asumir la función de núcleo del segundo sintagma (*differenti L1 > diferentes niveles de L1*); el adjetivo *interlinguistico* se recategoriza como sustantivo plural (*diverso livello interlinguistico > diferentes interlingüismos*):

EJEMPLO (8)	
TO	immigrati a Napoli con diverso *livello interlinguistico* e differenti *L1*.
GoogleT	inmigrantes extranjeros en Nápoles con diferentes *interlingüismos* y diferentes *niveles de L1*

2. *Traducción errónea por alucinación*. Aunque el debate en torno al fenómeno de las alucinaciones se haya intensificado con la llegada de los LLM, su estudio se remonta al desarrollo de los primeros sistemas de TAN (Koehn *et al.*, 2017; Lee *et al.*, 2018; Müller *et al.*, 2020; Raunak *et al.*, 2021). Diversos autores distinguen las alucinaciones que se generan con los sistemas de TAN de las que se generan con los LLM. En el primer caso, «... the output of the NMT system is often quite fluent [...] but completely unrelated to

the input» (Koehn *et al.*, 2017, p. 30). Las segundas son «cualitativamente distintas» (Guerreiro *et al.*, 2023, p. 1059), porque guardan cierta relación semántica con el contenido del texto origen (Sui *et al.*, 2024, p. 14 275). En este sentido, se ha propuesto en inglés el concepto de *confabulation* para referirse al mecanismo de narración generativa que, aunque ficcional, resulta coherente y verosímil (Smith *et al.*, 2023, 1; Sui *et al.*, 2024, p. 14 275)[17].

Los siguientes ejemplos son casos de *alucinación* encontrados en las traducciones de los TAN, en concreto, en GoogleT: el sintagma *cinco clases de anáforas* pasa al italiano como *cinque sintagmi anali*; *el uso de las comillas en los titulares de los periódicos* se traduce como *l'uso dei pulsanti nei titoli dei giornali;* el sintagma *las comillas dobles* se convierte en *le frasi a doppia intestazione*; por último, la oración *una comprensión mejor cuanto más próximos se hallan los elementos de la anáfora* se traduce como *una migliore comprensione quando gli elementi dell'anfora vengono allucinati*.

Muy distintas son las *alucinaciones* generadas por los LLM, más próximas al concepto de *confabulation*. El sistema inventa con coherencia y verosimilitud. En nuestro corpus, se documentan dos casos, ambos de Deepseek. El sistema añade al resumen (abstract) un título verosímil, pero falso (9) y una última frase conclusiva de corte evaluativo más propia de una reseña que de un resumen (10). Al estar relacionado con el contenido del TO, el error pasa desapercibido, aun cuando afecta gravemente a la fiabilidad de la traducción.

EJEMPLO (9)	
TO	Ø
Deepseek	Il linguaggio inclusivo come sfida traduttiva: strategie spagnole e italiane nell'era dell'intelligenza artificiale

EJEMPLO (10)	
TO	L'analisi mette in evidenza la grande versatilità testuale di questo dispositivo di coesione, che [...] si dimostra in grado di operare ad ampio raggio su diversi livelli di organizzazione del testo. Ø
Deepseek	El análisis revela la gran versatilidad textual de este mecanismo de cohesión, que [...] demuestra capacidad para operar a distintos niveles de organización textual. *Este estudio ofrece así una contribución significativa a la comprensión de los procesos de referencia y cohesión en el discurso escrito.*

3. *Traducción errónea por falta de enriquecimiento pragmático.* Es un error automático frecuente. Afecta a unidades sintagmáticas que necesitan ser explícitamente enriquecidas con información pragmática para garantizar su adecuada interpretación en el nuevo contexto. Véase el siguiente ejemplo:

[17] La noción de *alucinación* es un tecnicismo que procede del campo de la psicología y de la psiquiatría. Su aplicación al campo de la inteligencia artificial refleja la tendencia a antropomorfizar los sistemas. Lo mismo cabe decir de *confabulation*, un tecnicismo del inglés que podría traducirse al español como *fabulación.* En ámbito médico se documenta, sin embargo, *confabulación*, un calco semántico que recupera un sentido en desuso del término español.

EJEMPLO (11)	
TO	Tra i problemi di cui mi occupo c'è quello dei cosiddetti *nomina agentis*, cioè i nomi professionali, declinati al femminile (ministra, assessora, ecc.).

Al ser un artículo escrito en italiano, resulta innecesaria la especificación de la lengua sobre la cual se trabaja el fenómeno de los *nomina agentis*. La traducción, en cambio, debería hacerla explícita: *los denominados nomina agentis en italiano*. Este tipo de enriquecimiento contextual es una práctica habitual en la traducción humana, pero un reto en la automática. Ningún sistema resuelve el problema.

La misma tendencia se observa en la dirección opuesta. Véase el siguiente ejemplo (12).

EJEMPLO (12)	
TO	Este es el caso de los sintagmas preposicionales «en primer lugar», «en segundo lugar» y «por último», que pueden funcionar como adjuntos del verbo o bien, recategorizados como locuciones adverbiales y desplazados a la periferia, como ordenadores discursivos.

Todos los sistemas automáticos han traducido el sintagma preposicional (*in primo luogo*, *in secondo luogo*, *infine*) y ninguno de ellos ha enriquecido el sintagma con la especificación de la lengua objeto de estudio. Es un error que se documenta tanto en los TAN como en los LLM.

4. *Traducción errónea por incorrecta resolución anafórica.* La resolución de mecanismos fóricos representa un problema recurrente en los sistemas informáticos de traducción (Mitkov, 2014, pp. 818-845). La incorrecta recuperación de la anáfora puede afectar tanto a la precisión del texto, con el consiguiente falseamiento de lo dicho o explicatura, como a su fluidez, con el consiguiente aumento de los costes de procesamiento. Cuando afecta a la precisión, altera el contenido intencionalmente comunicado, bien introduciendo ambigüedades, bien devolviendo interpretaciones erróneas. Obsérvese el ejemplo (13).

EJEMPLO (13)	
TO	Pochi, invece, si sono avventurati sia nello studio dell'acquisizione di derivazione e composizione, sia nella comparazione, [...], delle categorie *(ma è davvero legittimo definirle tali?)* derivazionali e dei processi di composizione.
GoogleT	[...] de categorías derivacionales *(pero ¿es realmente legítimo definirlas como tales?)* y procesos de composición.
DeepL	[...] las categorías derivacionales *(¿pero es realmente legítimo llamarlas así?)* y los procesos de composición.
ChatGPT	[...] las categorías derivativas *(¿es realmente legítimo llamarlas así?)* y de los procesos de composición.
Deepseek	[...] categorías derivacionales *(¿es realmente apropiado definirlas así?)* y de los procesos de composición.

En (13), el determinante *tali* remite anafóricamente y sin ambigüedad a *categorie*. El desplazamiento del paréntesis propuesto por todos los sistemas en la traducción al español introduce, efectivamente, una mejora estilística que facilita la lectura, pero cambia la interpretación porque orienta de manera distinta la búsqueda del referente apuntando a *derivativas.* La traducción automática ha generado un problema que resulta, en cambio, de fácil resolución para el traductor humano, que hubiera optado por una anáfora léxica con repetición del lexema *categorías*: *categorías derivacionales (pero ¿es realmente apropiado definirlas categorías?)*.

En (14), la referencia en el TO se resuelve sin esfuerzo gracias a la capacidad anafórica del determinante demostrativo *esta*, que remite al elemento más próximo en el discurso anterior (*español*). Los sistemas neuronales falsean claramente la explicatura. Por su parte, los LLM generan una ambigüedad imposible de resolver: *nella stessa lingua* puede remitir al *español*, al *italiano* o incluso a otra lengua no mencionada explícitamente. La traducción humana sería *un'intervista in spagnolo sottotitolata in questa stessa lingua in due versioni*:

EJEMPLO (14)	
TO	visualizaron una entrevista en español subtitulada *en esta misma lengua* en dos versiones.
Google y Deepl	... un'intervista in spagnolo sottotitolata *in italiano* in due versioni.
ChatGPT	... un'intervista in spagnolo sottotitolata *nella medesima lingua* in due versioni.
Deepseek	... un'intervista in spagnolo sottotitolata *nella stessa lingua* in due versioni.

5.2.1.3. Entorno cultural

Esta categoría incluye, entre otros, aquellos errores que pueden herir la sensibilidad del destinatario o que pueden incluso resultar ofensivos, como los insultos o las blasfemias. La categoría incluye también el uso de un discurso no inclusivo y estereotipado, que puede manifestarse de forma más o menos oculta en el texto. El reconocimiento de estos elementos es fundamental para que la traducción, además de ser precisa, sea también ética y culturalmente adecuada.

La traducción en clave inclusiva es un reto para los sistemas de traducción automática, en particular para los sistemas neuronales. El ejemplo (15) muestra el distinto tratamiento que recibe la estrategia del desdoblamiento del texto original *los y las profesionales* en los sistemas neuronales y en los LLM. Los primeros eliminan las marcas, traduciendo con el masculino genérico *i professionisti,* que obliga a recuperar la interpretación de grupo mixto por vía inferencial. Los LLM, en cambio, garantizan la inclusividad, en línea con el significado intencionalmente comunicado por el TO. ChatGPT reformula con el epiceno *le figure professionali;* Deepseek recurre al desdoblamiento del artículo y de las desinencias con *i/le professionisti/e* (véase el ejemplo 15).

EJEMPLO (15)	
TO	la traducción en clave inclusiva supone un gran reto para *los y las profesionales* de la traducción.
Google y Deepl	la traduzione inclusiva rappresenta una sfida importante per *i professionisti* della traduzione.
ChatGPT	la traduzione in chiave inclusiva rappresenta una sfida considerevole per *le figure professionali* del settore.
Deepseek	la traduzione in ottica inclusiva costituisce una sfida rilevante per *i/le professionisti/e* del settore,

5.2.2. *Errores relacionados con la optimización del estímulo ostensivo en el TM*

5.2.2.1. Fluidez (convenciones lingüísticas)

Como se ha adelantado en §5.1.1. (tabla 5), los errores de fluidez son muy pocos y, en su mayoría, *menores.* Remiten a causas gramaticales, ortográficas y tipográficas, como, por ejemplo, el uso del artículo neutro *lo* en lugar de la forma masculina *el* en *recurrir a lo masculino llamado inclusivo* (Google); las erratas **encapsulamento* en lugar de *incapsulamento* o la ya señalada en §4.3.2. *ordinatori *discursivi* en lugar de *discorsivi*; o la sustitución de las comas de incisos por guiones breves en lugar de rayas, como prescribe la norma ortográfica del español, en el ejemplo (16).

EJEMPLO (16)	
TO	nella comparazione, in un'ampia prospettiva interlinguistica, [...].
Deepseek	como la comparación –desde una amplia perspectiva interlingüística–

Se documentan también errores de mayor impacto relacionados con la cohesión y la coherencia, como los que remiten a procedimientos anafóricos. Es el caso de los ejemplos (17) y (18). En (17), el sintagma *gli altri* debería flexionar en femenino para concordar con su referente *strategia*.

EJEMPLO (17)	
TO	La primera estrategia se orienta hacia la visibilización de la mujer desde el binarismo; *las demás* pretenden dar también visibilidad a las personas de género no binario
Google	La prima strategia è orientata alla visibilità delle donne da una prospettiva binaria; *Gli altri* mirano anche a dare visibilità alle persone di genere non binario

En (18), *los primeros* debería flexionar en femenino para concordar con su antecedente *señales de interacción*.

EJEMPLO (18)	
TO	A proposito della suddivisione [...] in *segnali di interazione* e segnali meta-testuali (Bazzanella, 1995), emerge una differenza tra gli informatori: *i primi* sono [...].
Deepseek	Respecto a la clasificación adoptada [...] en *señales de interacción* y señales metatextuales (Bazzanella, 1995)– se observan diferencias significativas entre los informantes: **los primeros* son...

Por lo que se refiere a la cohesión, GoogleT traduce literalmente el adverbio conjuntivo *donde* en lugar de traducir con *de ahí que*.

EJEMPLO (19)	
TO	Infine, si nota un uso più elevato di funzioni inferenziali negli apprendenti meno avanzati, *donde* l'ipotesi di stabilire una scala implicazionale
GoogleT	Finalmente, si se observa un mayor uso de funciones inferenciales en aprendices menos avanzados, *donde* la Hipótesis de establecer una escala implicacional

5.2.2.2. Estilo

Como explica Carrillo Guerrero (2005, pp. 135-136), el estilo puede entenderse «... como un aspecto o manifestación del discurso, [cuya realización] se hace en términos de variación: implica elección de léxico, estructuras, funciones... Esta variación está en función de todos los elementos que intervienen en la interacción comunicativa y de su contexto». El estilo se ve condicionado por las expectativas asociadas a la situación y a las distintas tradiciones discursivas; transgredirlas repercute directamente en la eficacia comunicativa (Carrillo Guerrero, 2005).

En esta categoría se documentan muy pocos errores. Son errores de registro y de naturalidad generados solo por los sistemas neuronales. En la dirección italiano-español, se documenta el uso inapropiado del tuteo *crear interferencias que te impidan hablar [...] el idioma que estás aprendiendo* (GoogleT). Hacia el italiano, el uso de unidades léxicas propias del registro coloquial, como *posto* en lugar de *posizione* para indicar una posición sintáctica (DeepL).

5.3. Elaboración discursiva con IA: corrección, reformulación estilística y lenguaje claro

Como hemos visto en los apartados anteriores, el análisis cualitativo ha permitido advertir tendencias de error propiamente automáticas. No menos relevante es el alto índice de variación estilística sin error que ha puesto de manifiesto la comparación entre sistemas. La variación se manifiesta en las dos direcciones y en todos los niveles de la lengua: léxico, sintáctico y macrosintáctico. Véanse los ejemplos (20) y (21). Indicamos en cursiva las unidades objeto de variación.

EJEMPLO (20)	
TO	*La presente contribución quiere contribuir a* la *comprensión* del *fenómeno*
GoogleT	*Questo contributo si propone di contribuire* alla comprensione del *fenomeno*
DeepL	*Il presente contributo si propone di contribuire* alla *comprensione* del *fenomeno*
ChatGPT	Il *presente contributo intende offrire* uno *spunto di riflessione* sul *fenomeno*
Deepseek	Il *presente contributo intende approfondire* la *questione*

EJEMPLO (21)	
TO	*Nello specifico*, le *ricerche di impronta tipologica* [...] in *ambito* morfologico lasciano ampiamente *sguarniti* due *interi settori* del componente morfologico
GoogleT	*En concreto, la investigación tipológica* [...] en el *campo* morfológico deja *en gran medida sin cubrir* dos *sectores enteros* del componente morfológico
DeepL	*Concretamente*, las *investigaciones tipológicas* [...] en el *ámbito* de la morfología dejan *en gran parte inexplorados* dos *sectores enteros* del componente morfológico
ChatGPT	*En concreto*, las *investigaciones de carácter tipológico* [...] en el *ámbito* morfológico dejan *en gran medida desatendidos* dos *sectores fundamentales* del componente morfológico
Deepseek	*En particular*, las *investigaciones* con [...] en el *ámbito* morfológico han dejado *prácticamente desatendidos* dos *ámbitos completos* del componente morfológico

La variación léxica se manifiesta también en la mayor o menor preferencia por el uso de préstamos directos del inglés: así, por ejemplo, el sintagma *fillers nella conversazione* utilizado en el TO italiano pasa como *rellenos en la conversación* (GoogleT y DeepL), *fillers conversacionales* (ChatGPT) y *elementos de relleno conversacional* (Deepseek); *tecnica del focus group* (TO) pasa como *técnica de grupo focal* (Google), *técnica del focus group* (ChatGPT) y *grupos focales* (Deepseek); el tecnicismo *code-switching* (TO) pasa como *cambio de código* (GoogleT), *code-switching* (Deepl, ChatGPT) y *cambio de código (code-switching)* (Deepseek).

En el contexto de este alto índice de variación, los LLM destacan por la elección de variantes léxicas más acertadas desde el punto de vista del registro y de la terminología. Así, por ejemplo, GoogleT y DeepL traducen *molte volte* como *muchas veces*, mientras que ChatGTP y Deepseek proponen, respectivamente, *en numerosas ocasiones* y *en muchas ocasiones*. Del mismo modo, el tecnicismo *lenguas de especialidad* (TO) se traduce con las variantes *lingue settoriali* (ChatGPT) e *linguaggi specialistici* (Deepseek), ambas bien consolidadas por el uso, frente a la más literal y menos frecuente *linguaggi specializzati* (GoogleT y DeepL).

Por otro lado, en el marco de la teoría de la pertinencia, el estilo es una elección de formulación que se interpreta en función de los efectos cognitivos y del esfuerzo de procesamiento para alcanzarlos (Sperber y Wilson 1994, p. 248) y, por ello, está directamente relacionado con la claridad. La dimensión estilística así entendida resulta particularmente importante en la comunicación especializada y, por ende, en la traducción especializada. Un estilo claro es sinónimo de pertinencia y eficacia comunicativas. En este sentido, en el corpus de traducciones se han identificado tres tendencias automáticas meliorativas del texto original:

1. *Corrección de errores materiales*: se refiere a la capacidad de identificar y corregir erratas, basándose en el cotexto. Son ejemplos como los siguientes: en el ejemplo (22) figura **abbandonanti* en lugar de *abbondanti*. Los LLM resuelven el problema. Las versiones de GoogleT y DeepL carecen completamente de sentido. Obsérvese el fuerte impacto que tiene la ausencia de corrección de la errata en términos de coherencia:

EJEMPLO (22)	
TO	… emerge una differenza tra gli informatori: i primi sono **abbandonanti* nei dati di apprendenti avanzati
GoogleT	… surge una diferencia entre los informantes: los primeros son *abandoners* en los datos de estudiantes avanzados
DeepL	… surge una diferencia entre los informantes: los primeros se *abandonan* en los datos de los alumnos avanzados
ChatGPT	… se observa una diferencia entre los informantes: las primeras son *abundantes* en los datos de aprendices avanzados
Deepseek	… se observan diferencias significativas entre los informantes: los primeros son *abundantes* en los datos de aprendientes avanzados

En (23), ChatGPT, Deepseek y DeepL corrigen la falta de espacio entre el sustantivo *antecedente* y la conjunción. GoogleT no advierte la errata y genera una traducción carente de sentido, dado que la preposición *tra* necesita de una estructura bimembre.

EJEMPLO (23)	
TO	En cada una de ellas se ha manipulado la distancia entre el *antecedentey* elemento fórico […].
GoogleT	In ogni testo, la distanza tra l'elemento forico precedente è stata manipolata
DeepL	In ognuno di essi è stata manipolata la distanza tra l'antecedente e l'elemento fonico
ChatGPT	Per ciascuna categoria, è stata manipolata la distanza tra l'antecedente e l'elemento anaforico
Deepseek	In ciascuna di esse è stata manipolata la distanza tra l'antecedente e l'elemento forico

2. *Resolución de ambigüedades.* En el ejemplo (24), la locución preposicional *a partir de* del TO permite una doble lectura: *a*) el corpus está compuesto solo por unidades de la primera edición y *b*) el corpus es dinámico y se enriquece con unidades de todas las ediciones. La traducción de GoogleT y Deepseek es literal y correcta; DeepL y ChatGTP mejoran el resultado porque resuelven la ambigüedad.

EJEMPLO (24)	
TO	se ha decidido compilar un corpus de unidades fraseológicas *a partir de* la primera edición del *Diccionario de uso del español* para luego contrastarlo tanto con las posteriores ediciones del mismo DUE […]
GoogleT	si è deciso di compilare un corpus di sintagmi fraseologici unificati *a partire dalla prima* edizione del *Dizionario dell'uso dello spagnolo* per confrontarlo con entrambe le edizioni successive dello stesso DUE […]
DeepL	si è deciso di compilare un corpus di unità fraseologiche *tratte dalla prima* edizione del *Diccionario de uso del español* e di confrontarlo poi sia con le edizioni successive del DUE […]
ChatGPT	si è deciso di compilare un corpus di unità fraseologiche *tratte dalla* prima edizione del *Diccionario de uso del español*, per poi confrontarlo sia con le edizioni successive dello stesso DUE […]
Deepseek	si è deciso di compilare un corpus di unità fraseologiche *a partire dalla* prima edizione del *Diccionario de uso del español* per poi confrontarlo sia con le successive edizioni dello stesso DUE […]

3. *Reformulación y mejora estilística*. Los LLM destacan por su contribución a la mejora estilística del texto en términos de claridad: mejor accesibilidad cognitiva, menores costes de procesamiento, mayor fluidez y mayor pertinencia. Esto se verifica incluso en ausencia de instrucciones específicas en este sentido (recordamos que los textos se han generado a partir de un *prompt básico*). Nótese la adición del marcador discursivo *ad esempio* en (25) y la explicitación del pronombre demostrativo anafórico *essi, esse* en (26). Ambas operaciones contribuyen a la mayor claridad y legibilidad del texto.

	EJEMPLO (25)
TO	En este movimiento argumentativo, el denominado complemento de proporción (para el precio, la ubicación no está ni tan mal)
GoogleT DeepL	il cosiddetto complemento di proporzione (per il prezzo, la località non è poi così male)
ChatGPT Deepseek	il cosiddetto complemento di proporzione (*ad esempio*: *per il prezzo, la posizione non è poi così male*)

	EJEMPLO (26)
TO	Con esta segunda función, indican [...].
GoogleT DeepL	Con questa seconda funzione, *Ø indicano* [...].
ChatGPT	In quest'ultima funzione, *essi* [gli ordinatori discorsivi] segnalano [...].
Deepseek	In questa seconda funzione, *esse* [le espressioni preposizionali] segnalano [...]

En (27), la reformulación propuesta por ChatGPT y Deepseek introduce una mejora significativa del TO en términos de claridad y accesibilidad al mensaje, en línea con los principios del *lenguaje claro*.

	EJEMPLO (27)
TO	la funzione meta-testuale mira a rendere l'enunciato quanto più coerente e coeso possibile, *il che non è trascurabile, ma può esserlo* per quegli apprendenti con livello interlinguistico non elevato
Google	La función metatextual pretende hacer el enunciado lo más coherente y cohesivo posible, *lo que no es despreciable, pero puede serlo* para aquellos aprendices con un nivel interlingüístico bajo
DeepL	la función metatextual tiene como objetivo que el enunciado sea lo más coherente y cohesionado posible, *lo cual no es baladí*, *pero puede serlo* para aquellos alumnos con un nivel interlingüístico bajo
ChatGPT	la función metatextual busca dotar al enunciado de la mayor coherencia y cohesión posible, *lo cual no es desdeñable*, pero *puede pasar a un segundo plano* en aquellos aprendices con un nivel interlingüístico bajo
Deepseek	la función metatextual, por su parte, busca dotar al enunciado de mayor coherencia y cohesión, aspecto que *-aunque relevante- puede resultar secundario* para aquellos aprendientes con niveles interlingüísticos más bajos

La formulación original es gramaticalmente correcta, pero muy costosa de procesar y estilísticamente poco acertada. Las traducciones de Google y DeepL mantienen el estilo original, sin aportar mejoras a la comprensibilidad del texto. En cambio, ChatGPT y Deepseek someten el texto a diversas operaciones de elaboración discursiva de base léxica, sintáctica y macrosintáctica. Las traducciones resultan más claras: mejora el nivel de informatividad del texto y los costes de procesamiento se reducen.

En definitiva, los LLM introducen mejoras estilísticas que se traducen en un nivel mayor de accesibilidad cognitiva y de claridad informativa. Es un aspecto relevante que no puede no tenerse en cuenta en la evaluación del grado de eficacia de los sistemas. De hecho, cada vez más se están proponiendo enfoques de traducción de textos complejos que consideran los principios del lenguaje claro en distintos ámbitos de especialidad, entre ellos, el académico (Marazzato Sparano, 2018; RAE / ASALE, 2024; Retegui y Rocca, 2024)[18].

6. CONCLUSIONES

La irrupción de los LLM en la traducción está transformando de forma radical la práctica profesional; sin embargo, es todavía muy escasa la investigación orientada a la evaluación de su grado de eficacia en comparación con la de los sistemas neuronales, en particular para la combinación lingüística español-italiano / italiano-español. Con el presente estudio se ha querido contribuir a colmar este vacío, focalizando la atención en el discurso académico a partir de un corpus de creación propia constituido por diez resúmenes de contribuciones científicas del italiano al español y otros diez del español al italiano, con un total de 1685 y 1645 palabras, respectivamente. El tamaño reducido del corpus no permite sacar conclusiones estadísticamente generalizables, pero sí de interés cualitativo. La anotación de los errores ha sido completamente humana y se ha utilizado la métrica MQM (Lommel *et al.*, 2014). Desde un punto de vista teórico, se ha adoptado la teoría de la relevancia (Sperber y Wilson, 1986; Wilson y Sperber, 2004) para dar cuenta de la gravedad de los errores y explicar los problemas que se verifican durante el proceso de traducción. Además, ha servido de base para proponer, desde supuestos cognitivos, una escala de gravedad más ajustada al campo discursivo objeto de estudio.

Se han planteado dos objetivos principales: en primer lugar, evaluar empírica y comparativamente la calidad de los resultados y el grado de eficacia de dos sistemas de TAN (DeepL y Google Translate) frente a dos LLM (ChatGPT y Deepseek) y, en segundo lugar, identificar, clasificar y describir cualitativamente los tipos de errores generados, así como individuar fenómenos y patrones específicos de traducción automática que no se verifican en la humana. El análisis se ha realizado desde el punto de vista tanto cuantitativo como cualitativo. Los resultados confirman las hipótesis de partida.

---

[18] Véase, en este sentido, Parlamento Europeo (s. f.). «Evolución de la profesión», en https://www.europarl.europa.eu/translation/es/translation-at-the-european-parliament/how-the-profession-is-evolving (consultado el 16 de octubre de 2025). Más dirigido a la redacción asistida de textos de especialidad en español, véase el proyecto Artext dirigido por Iria da Cunha, de la Universidad Nacional de Educación a Distancia: https://sistema-artext.com/es/.

REVISTA DE FILOLOGÍA, 52; 2026, PP. 379-416

En primer lugar, los sistemas generativos han demostrado un mejor rendimiento que los neuronales. La puntuación MQM coloca a los LLM por delante de TAN, con un número menor de errores graves (mayores y críticos) en ambas direcciones de traducción.

En segundo lugar, la mayor parte de los errores está relacionada con la reconstrucción de la explicatura de nivel inferior y, concretamente, con las categorías de *Terminología* y *Precisión*. La existencia de un número significativo de errores en estas dos categorías ha permitido, además, identificar siete tendencias automáticas de error que son transversales a todos los sistemas y que responden claramente a mecanismos de razonamiento no humano de base algorítmica. Por lo que se refiere a la *Terminología*, se ha observado la tendencia a traducir según criterios de semejanza puramente formal, a alejarse de la literalidad incluso en aquellos casos en los que la opción más literal hubiera sido la más adecuada, y a recurrir a tecnicismos pertenecientes a otros ámbitos de especialidad. Por lo que se refiere a la *Precisión*, las tendencias observadas son cuatro: en concreto, *traducción errónea* por reanálisis composicional, por *alucinación* y *fabulación*, por falta de enriquecimiento pragmático y por incorrecta resolución anafórica. En esta línea, serían deseables ulteriores estudios con vistas a confirmar o precisar las tendencias identificadas, o añadir otras nuevas.

En tercer lugar, se ha afrontado el análisis de la dimensión estilística del texto traducido. La comparación entre sistemas ha puesto de manifiesto un alto índice de variación estilística que afecta a todos los niveles de la lengua (léxico, sintáctico y macrosintáctico) en ambas direcciones. Los LLM se han revelado mejores en la elección de la variante más adecuada en términos de registro. En relación con la dimensión variacional se han advertido, además, tres tendencias meliorativas de elaboración discursiva con respecto al texto original: corrección de errores materiales, resolución de ambigüedades y reformulación y mejora estilística. En esta última función, como se había supuesto, los LLM destacan por su capacidad para activar procesos complejos de elaboración discursiva con impacto en el estilo global del texto meta en términos de claridad y de accesibilidad al mensaje, en línea con los supuestos del denominado *lenguaje claro*. Es un aspecto relevante que no puede pasarse por alto en la evaluación del grado de eficacia de los sistemas.

En cuarto lugar, desde el punto de vista del par de lenguas considerado, la puntuación MQM obtenida por los sistemas es más elevada en la dirección italiano-español. La diferencia resulta más marcada en los sistemas neuronales. Cabe suponer que esto se deba a la condición de lengua global del español, con una mayor presencia en la red y como lengua meta de traducción (Instituto Cervantes, 2024, pp. 77-78 y 82). Se necesitarían ulteriores investigaciones con corpus de mayor alcance para llegar a conclusiones generalizables en este sentido.

Por último, los resultados de la investigación confirman la centralidad del traductor humano, figura necesaria y no sustituible en cuanto garante de la calidad última del texto. En el marco contemporáneo de la *traducción aumentada* (DePalma, 2017), equiparse con nuevos saberes y competencias tecnológicas, así como comprender en profundidad los fenómenos propios de la traducción automática, constituye un requisito indispensable para el desarrollo de la competencia traductora.

El presente estudio esperamos pueda servir de punto de partida para futuras investigaciones sobre otros ámbitos de especialización o con corpus de mayor alcance, o sobre aspectos concretos de interés emergente, como el impacto del diseño del *prompt* en la calidad del resultado.

Recibido: 21.10.2025; aceptado: 11.02.2026.

BIBLIOGRAFÍA

ALAMMAR, Jay (27 de junio de 2018). *The Illustrated Transformer.* GitHub. https://jalammar.github.io/illustrated-transformer/. Consultado el 7 de septiembre de 2025.

BALASHOV, Yuri (2025). Translation in the Wild. arXiv. https://doi.org/10.48550/arXiv.2505.23548.

BRIAKOU, Eleftheria, CHERRY, Colin y FOSTER, George (2023). Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1. https://doi.org/10.18653/v1/2023.acl-long.524.

BROWN, Tom *et al.* (2020). *Language Models are Few-Shot Learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165.

CALVI, Maria Vittoria, BORDONABA ZABALZA, Cristina, MAPELLI, Giovanna y SANTOS LÓPEZ, Javier (Eds.). (2011). *Las lenguas de especialidad en español* (1.ª ed.). Carocci editore.

CARRILLO GUERRERO, Lázaro (2005). Marco comunicativo del estilo en el uso de la lengua. *Ámbitos. Revista Internacional de Comunicación.* 13-14, 135-153. https://doi.org/10.12795/Ambitos.2005.i13-14.09.

CID-LEAL, Pilar, ESPÍN-GARCÍA, María del Carmen y PRESAS, Marisa (2019). Traducción automática y posedición: Perfiles y competencias en los programas de formación de traductores. *MonTi: Monografías de Traducción e Interpretación*, 11, 187-214.

CRISTIANINI, Nello (2024). *Machina sapiens. L'algoritmo che ci ha rubato il segreto della conoscenza.* Il Mulino.

DA CUNHA, Iria (2024). Inteligencia artificial y lenguaje claro en español. En Alejandro Rafael Retegui y Fernando Bernabé Roca (Eds.). *Lenguaje claro en Iberoamérica. Principios y prácticas* (pp. 417-444). Thomson Reuters-La Ley Argentina.

DE PALMA, Donald A. (10 de febrero 2021). *Augmenting Human Translator Performance*. CSA Research. Recuperado de https://csa-research.com/l/blog/article/Augmenting-Human-Translator-Performance.

DE PALMA, Donald A. (15 de febrero de 2017). *Augmented Translation Powers up Language Services*. CSA Research. Recuperado de https://csa-research.com/l/blog/article/Augmented-Translation-Powers-up-Language-Services.

DELISLE, Jean (1993). *La traduction raisonnée: Manuel d'initiation à la traduction professionnelle anglais français: méthode par objectifs d'apprentissage* (1.ª reimp.). Presses de l'Univ. d'Ottawa.

DURO MORENO, Miguel (2012). Entornos y determinación de la traducción jurídica inglés-español. *Analecta Malacitana,* II (Anejo LXXXVI), 301-327.

ERVAS, Francesca (2010). A naturalistic explanation of communication across cultures. *Trans. Zeitschrift fürKulturwissenschaften, 17*(8.14). https://www.inst.at/trans/17Nr/8-14/8-14_ervas.htm.

FORCADA, Mikel L. (2017). Making sense of neural machine translation. *Translation Spaces*, *6*(2), 291-309. https://doi.org/10.1075/ts.6.2.06for.

GLADKOFF, Serge, SOROKINA, Irina, HAN, Lifeng y ALEKSEEVA, Alexandra (2022). Measuring uncertainty in translation quality evaluation (TQE). *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (1454-1461). European Language Resources Association.

GÓMEZ-RODRÍGUEZ, Carlos (2025). Grandes modelos de lenguaje: ¿de la predicción de palabras a la comprensión? En Amparo Alonso Betanzos, Daniel Peña y Pilar Poncela (Eds.), *La inteligencia artificial hoy y sus aplicaciones con big data* (pp. 73-98). Funcas.

Grice, Herbert Paul (1975). Logic and conversation. En Herbert Paul Grice (1989), *Studies in the way of words* (pp. 22-40). Harvard University Press.

Guerreiro, Nuno M., Alves, Duarte M., Waldendorf, Jonas, Haddow, Barry, Birch, Alexandra, Colombo, Pierre y Martins, André F. T. (2023). Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11, 1500-1517. https://doi.org/10.1162/tacl_a_00615.

Gutt, Ernst-August (2010). *Translation and Relevance: Cognition and Context* (2.ª ed). Taylor and Francis.

Hendy, Amr, Abdelrehim, Mohamed, Sharaf, Amr, Raunak, Vikas, Gabr, Mohamed, Matsushita, Hitokazu, Kim, Young Jin, Afify, Mohamed y Awadalla, Hany Hassan (2023). *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation*. arXiv. https://doi.org/10.48550/arXiv.2302.09210.

Holdsworth, Jim y Scapicchio, Mark (17 de junio de 2024): *¿Qué es el deep learning?* IBM. Recuperado de https://www.ibm.com/es-es/think/topics/deep-learning.

Hurtado Albir, Amparo (2007). *Traducción y traductología: Introducción a la traductología* (3.ª ed). Cátedra.

Instituto Cervantes (2024). *El español en el mundo: Anuario del Instituto Cervantes, 2024*. Instituto Cervantes. https://cvc.cervantes.es/lengua/anuario/anuario_24/default.htm

Kalchbrenner, Nal y Blunsom, Phil (2013). Recurrent Continuous Translation Models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700-1709. https://doi.org/10.18653/v1/D13-1176

Kocmi, Tom, Avramidis, Eleftherios, Bawden, Rachel, Bojar, Ondřej, Dvorkovich, Anton, Federmann, Christian, Fishel, Mark, Freitag, Markus, Gowda, Thamme, Grundkiewicz, Roman, Haddow, Barry, Karpinska, Marzena, Koehn, Philipp, Marie, Benjamin, Monz, Christof, Murray, Kenton, Nagata, Masaaki, Popel, Martin, Popović, Maja, Shmatova, Mariya ... Zouhar, Vilém (2024). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. *Proceedings of the Ninth Conference on Machine Translation*, 1-46. https://doi.org/10.18653/v1/2024.wmt-1.1

Koehn, Philipp (2020). *Neural machine translation*. Cambridge University Press. https://doi.org/10.1017/9781108608480.

Koehn, Philipp y Knowles, Rebecca (2017). Six Challenges for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28-39. https://doi.org/10.18653/v1/W17-3204.

Kornacki, Michal y Pietrzak, Paulina (2025). *Hybrid workflows in translation: Integrating GenAI into translator training*. Routledge.

Lee, Katherine, Firat, Orhan, Agarwal, Ashish, Fannjiang, Clara, y Sussillo, David (2018). *Hallucinations in Neural Machine Translation*. ICLR 2019 Conference, recuperado de OpenReview.net: https://openreview.net/forum?id=SkxJ-309FQ.

Lommel, Arle, Gladkoff, Serge, Melby, Alan, Wright, Sue Ellen, Strandvik, Ingemar, Gasova, Katerina, Vaasa, Angelica, Benzo, Andy, Marazzato Sparano, Romina, Foresi, Monica, Innis, Johani, Han, Lifeng y Nenadic, Goran (2024). *The Multi-Range Theory of Translation Quality Measurement: MQM scoring models and Statistical Quality Control*. arXiv. https://doi.org/10.48550/arXiv.2405.16969.

Lommel, Arle, Uszkoreit, Hans y Burchardt, Aljoscha (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica Tecnologies de La Traducció*, 12, 455-463. https://doi.org/10.5565/rev/tradumatica.77.

Lozano Zahonero, María (2021). Traducción y tecnología. En G.B. Félix San Vicente (Ed.), *Lengua española para traducir e interpretar* (pp. 225-243). CLUEB.

Mapelli, Giovanna (2009). El resumen o abstract: género, estructura y rasgos discursivos. En Maria Vittoria Calvi, Cristina Bordonaba Zabalza, Giovanna Mapelli y Javier Santos López (Eds.), *Las lenguas de especialidad en español* (pp. 101-121). Carocci.

Marazzato Sparano, Romina (2018). Lenguaje claro, traducción e idiosincrasias del idioma: aportes para la comprensión lectora. *Orientación y sociedad*, *18*(2), 163-177.

Minervini, Rosaria (2021). La traducción automática del género (español-italiano): Análisis de ejemplos traducidos con DeepL y Google Traductor. *Rivista internazionale di tecnica della traduzione*, 23, 105-127. EUT Edizioni Università di Trieste. https://doi.org/10.13137/2421-6763/33237.

Minervini, Rosaria (2023). La traducción automática español-italiano del turismo enogastronómico: Un estudio de caso. *Cuadernos de Lingüística Hispánica*, 42, 1-20. https://doi.org/10.19053/0121053X.n42.2023.16000

Mitchell, Melanie (2022). *L'intelligenza artificiale*. Einaudi.

Mitkov, Ruslan (Ed.). (2014). *The Oxford handbook of computational linguistics* (1.ª ed. en línea). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.001.0001.

Montero Martínez, Silvia, Blanchar Faber Benítez, Pamela, y Buendía Castro, Miriam (2011). *Terminología para traductores e intérpretes* (2.ª ed.). Ediciones Tragacanto.

Moorkens, Joss, Castilho, Sheila, Gaspari, Federico y Doherty, Stephen (Eds.). (2018). *Translation quality assessment: From principles to practice*. Springer.

MQM Council (2025). *MQM (Multidimensional Quality Metrics)*. TheMQM.org. Recuperado de https://themqm.org/.

Müller, Mathias, Rios, Annette, y Sennrich, Rico (2020). Domain robustness in neural machine translation. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*, 1, 151-164.

Nord, Christiane (1994). Translation as a process of linguistic and cultural adaptation. En Cay Dollerup y Annette Lindegaard (Eds.), *Teaching Translation and Interpreting 2: Insights, aims and visions. Papers from the Second Language International Conference Elsinore, 1993* (pp. 59-68). John Benjamins Publishing Company.

O'Donnell, Michael James (2008). *The UAM CorpusTool: Software for corpus annotation and exploration*. *Actas del XXVI Congreso de AESLA*.

O'Donnell, Michael James (2021). The UAM CorpusTool: Software for corpus annotation and exploration. En Carmen M.ª Bretones Callejas *et al.* (Eds.), *Applied linguistics now: Understanding language and mind/La Lingüística Aplicada Actual: Comprendiendo el lenguaje y la mente* (pp. 1433-1447). Universidad de Almería.

Pantcheva, Marina (29 de abril de 2025). *How do we train LLMs for machine translation?* AI Localization Think Tank. https://www.ailocthinktank.com/post/how-do-we-train-llms-for-machine-translation.

Parlamento Europeo (s. f.). Evolución de la profesión. Recuperado de https://www.europarl.europa.eu/translation/es/translation-at-the-european-parliament/how-the-profession-is-evolving.

Portolés Lázaro, José (2002). Marcadores discursivos y traducción. En Joaquín García Palacios y M.ª Teresa Fuentes Morán (Eds.), *Texto, terminología y traducción* (pp. 145-167). Almar.

Portolés Lázaro, José (2016). Marcadores del discurso. En Javier Gutiérrez-Rexach (Ed.), *Enciclopedia de lingüística hispánica,* volumen 1 (pp. 689-699). Routledge.

Raunak, Vikas, Menezes, Arul y Junczys-Dowmunt, Marcin (2021). The Curious Case of Hallucinations in Neural Machine Translation. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1172-1183. https://doi.org/10.18653/v1/2021.naacl-main.92.

Real Academia Española y Asociación de Academias de la Lengua Española (2024). *Guía panhispánica de lenguaje claro y accesible*. Espasa-Calpe.

Rehm, Georg (2025, marzo 19). *The European Language Data Space: Overview* [Presentación en conferencia]. LDS Launch Conference, Villers-Cotterêts, Francia. https://language-data-space.ec.europa.eu/system/files/2025-03/01_Overview.pdf.

Retegui, Alejandro Rafael y Rocca, Fernando Bernabé (2024). *Lenguaje claro en Iberoamérica: Principios y prácticas (1.ª ed.)*. La Ley.

Riina, Nicholas, Patlolla, Likhitha, Hernandez Joya, Camilo, Bautista, Roger, Olivar-Villanueva, Melissa y Kumar, Anish (2024). An evaluation of English to Spanish medical translation by large language models. En Marianna Martindale, Janice Campbell, Konstantin Savenkov y Shivali Goel (Eds.), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)* (pp. 222-236). https://aclanthology.org/2024.amta-presentations.15/.

Sánchez Ramos, María del Mar y Rico Pérez, Celia (2020). *Traducción automática: Conceptos clave, procesos de evaluación y técnicas de posedición.* Editorial Comares.

Sharma, Sonali, Diwakar, Manoj, Singh, Prabhishek, Singh, Vijendra, Kadry, Seifedine y Kim, Jungeun (2023). Machine Translation Systems Based on Classical-Statistical-Deep-Learning Approaches. *Electronics*, *12*(7), 1716. https://doi.org/10.3390/electronics12071716.

Smith, Andrew L., Greaves, Felix y Panch, Trishan (2023). Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. *PLOS Digital Health*, *2*(11), 1-3. https://doi.org/10.1371/journal.pdig.0000388.

Specia, Lucia y Wilks, Yorick (2016). *Machine Translation. The Oxford Handbook of Computational Linguistics* (pp. 817-870). https://doi.org/10.1093/oxfordhb/9780199573691.013.26.

Sperber, Dan (Ed.). (2000). *Metarepresentation. A Multidisciplinary Perspective*. Oxford University Press.

Sperber, Dan y Wilson, Deirdre (1994). *La relevancia: Comunicación y procesos cognitivos*. (Traducción de E. Leonetti). Visor Dis. (Trabajo original publicado en 1986).

Sperber, Dan, y Wilson, Deirdre (1986). *Relevance: Communication and cognition*. Blackwell.

Sui, Peiqi, Duede, Eamon, Wu, Sophie y So, Richard (2024). Confabulation: The Surprising Value of Large Language Model Hallucinations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1, 14274-14284. https://doi.org/10.18653/v1/2024.acl-long.770.

Torrejón, Enrique y Rico, Celia (2012). Habilidades y perfil del nuevo rol del traductor como poseditor de traducción automática. *Tradumàtica tecnologies de la traducció*, 10, 166-178. https://doi.org/10.5565/rev/tradumatica.18.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jamob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz y Polosukhin, Illia (2017). Attention Is All You Need. *Advances in neural information processing systems*, 30. https://doi.org/10.48550/arXiv.1706.03762.

Wilson, Deirdre (2000). Metarepresentation in Linguistic Communication. En Dan Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (pp. 441-448). Oxford University Press.

Wilson, Deirdre y Sperber, Dan (2004). Relevance theory. En Laurence R. Horn y Gregory Ward (Eds.), *The handbook of pragmatics* (pp. 607-632). Blackwell Publishing.

Wilson, Deirdre y Sperber, Dan (2012). *Meaning and Relevance* (1.ª ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139028370.