Gutiérrez Rubio, Enrique

Aproximación al análisis de la calidad de la traducción automática desde la perspectiva de la evaluación basada en el usuario : el caso de DeepL, Google Translate y ChatGPT en la combinación checo-español

Études romanes de Brno. 2024, vol. 45, iss. 4, pp. 65-86

ISSN 2336-4416 (online)

Stable URL (DOI): https://doi.org/10.5817/ERB2024-4-4
Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.81314

ÉTUDES ROMANES DE BRNO 45/2024/4
HTTPS://DOI.ORG/10.5817/ERB2024-4-4

Aproximación al análisis de la calidad de la traducción automática desde la perspectiva de la evaluación basada en el usuario: el caso de DeepL, Google Translate y ChatGPT en la combinación checo-español

An Approach to User-Centered Translation Quality Assessment of Machine Translation Output: the Case of DeepL, Google Translate, and ChatGPT in Czech-to-Spanish Translation Outputs

Enrique Gutiérrez Rubio [enrique.gutierrez@upol.cz]
Univerzita Palackého v Olomouci, República Checa

Resumen
La generalización del uso de las aplicaciones gratuitas de traducción automática basadas en redes neuronales (NMT) demanda un mayor esfuerzo por parte de la comunidad científica por evaluar su calidad. En este artículo se presenta el estado de la cuestión, así como los resultados de un análisis piloto que trata de sacar a la luz el grado de satisfacción de los potenciales usuarios de estas traducciones en función de tres variables: fluidez, corrección gramatical y usabilidad. Con este fin, se realizó un experimento en el que veinte anotadores nativos de español evaluaron mediante una escala de valoración Likert las traducciones generadas por humanos profesionales y por las aplicaciones DeepL, Google Translate y ChatGPT de tres textos checos de diversa tipología (uno técnico, uno de *marketing* y uno literario). Los resultados muestran que, a pesar de que las traducciones humanas son las mejores valoradas, existe un elevado grado de satisfacción por parte de los usuarios respecto a las traducciones generadas por los sistemas NMT diseñados específicamente para este fin (DeepL y Google Translate) y, muy especialmente, en términos de fluidez y usabilidad.

Palabras clave
Traducción automática neuronal; evaluación de la calidad; traducción checo-español; DeepL; Google Translate; ChatGPT

Abstract
The widespread use of free Neural Machine Translation (NMT) systems requires a greater effort on the part of the scientific community to evaluate their quality. This article presents the state of the art and the results of a pilot analysis aimed at revealing the level of satisfaction of potential users of these translations in terms of three variables: fluency, grammar, and usability. To this end, an experiment was carried out in which twenty native Spanish annotators evaluated, using a Likert rating scale, the translations generated by human professionals and by the applications DeepL, Google Translate, and ChatGPT of three Czech

texts of different types (one technical, one marketing and one literary). The results show that although human translations are the best rated, there is a high degree of user satisfaction with the translations generated by NMT systems specifically designed for this purpose (DeepL and Google Translate), especially in terms of fluency and usability.

Keywords
Neural Machine Translation; translation quality assessment; Czech-Spanish translation; DeepL; Google Translate; ChatGPT

RECIBIDO 2024-03-01; **ACEPTADO** 2024-06-30

La financiación para esta investigación ha sido otorgada a la Universidad Palacký de Olomouc por el Ministerio de Educación, Juventud y Deporte de la República Checa (IGA_ FF_2023_032).

1. Introducción

Los programas con tecnología de traducción automática neuronal (*Neural Machine Translation* en inglés – NMT) ya son una realidad de nuestro día a día, de modo que, en apenas unos años, su uso ha dejado de ser experimental para generalizarse entre particulares y profesionales de todos los ámbitos con millones de usuarios a diario (Way 2018: 159; Hassan *et al.* 2018).[1] Las aplicaciones DeepL, Google Translate o Microsoft Translator, entre otras, emplean una arquitectura de redes neuronales artificiales que entrenan con miles de millones de parámetros. Autores como Hassan *et al.* (2018) afirmaban hace ya más de un lustro que, al menos en la combinación chino-inglés y en el ámbito de las noticias publicadas en Internet, las traducciones automáticas de Microsoft Translator eran tan potentes que igualaban en calidad a las realizadas por los traductores profesionales. Sin embargo, como señalan Fomicheva *et al.* (2020), la calidad de la traducción resultante depende de varios factores, como los pares de lenguas y el género/tipología del texto original. En este artículo se va a presentar brevemente la problemática de la tecnología de traducción NMT, así como la cuestión de la evaluación de las traducciones generadas por estas aplicaciones basadas en inteligencia artificial. A continuación, se aportarán los datos de un estudio piloto que analiza la calidad de varias traducciones realizadas por traductores profesionales y traductores automáticos sin proceso de posedición en diversos géneros (literatura, textos de *marketing* y textos técnicos) en la combinación no anglocéntrica checo-español. Esta investigación no se centra en la evaluación de la calidad del texto meta (TM) a través de un proceso comparativo con el texto origen (TO), sino en la valoración de la experiencia de los usuarios o lectores potenciales de las traducciones, una aproximación teórica que, en nuestra opinión, no ha recibido la atención que se merece y, muy especialmente, en combinaciones no anglocéntricas como sería el caso checo-español.

1 En este sentido, DeepL afirma que más de mil millones de personas han usado sus servicios (https://www.deepl.com/files/press/companyProfile_EN.pdf; consultado 27.5.2024).

2. Tecnología de traducción automática neuronal

Como ya hemos adelantado en la Introducción, los traductores automáticos basados en redes neuronales o NMT[2] se han convertido en una realidad de nuestro día a día. En algunos casos, como Google Translate o la versión básica de DeepL, se trata de herramientas gratuitas que, además, en los últimos años han mostrado una mejora sustancial de los servicios ofrecidos en términos de precisión y eficacia (Wang *et al.* 2023), lo que facilita que su uso se haya generalizado rápidamente. Tanto es así que muchas empresas, en lugar de recurrir a agencias de traducción, abaratan costes utilizando estos traductores automáticos, ya sea con un proceso de posedición (Castilho y O'Brien 2016) que busque mejorar el resultado final y la fidelidad de la información o sin él.[3] Sin embargo, estas tecnologías de traducción basadas en inteligencia artificial (IA) se hallan aún en fase de desarrollo y son muchos los especialistas que advierten de que sus resultados no son aún comparables con los producidos por los especialistas humanos. Así lo señalan los propios traductores profesionales de acuerdo con una encuesta realizada por Zaretskaya *et al.* (2015): la mayoría reconoce que la calidad de las traducciones automáticas es muy baja, de modo que los profesionales que emplean traductores automáticos para sus propios encargos confiesan tener que corregir después entre el 30 % y el 90 % de sus resultados. También la mayoría de los encuestados por Gaspari *et al.* (2015), concretamente el 52 %, afirma que la calidad de los sistemas MT es *pobre* o *baja,* frente al 11 % que la considera *alta* o *excelente*. Sin embargo, en ambos casos se trata de datos recogidos hace diez años, periodo en que esta tecnología con base en la inteligencia artificial ha mejorado de manera sustancial, especialmente gracias al desarrollo de los traductores automáticos basados en redes neuronales.[4] Tanto es así que, como ya ha sido aquí comentado, en la traducción automática chino-inglés de noticias publicadas en Internet, se habría alcanzando la *paridad humana*: "Our evaluation found that our [Microsoft machine translation] system has reached parity with professional human translations on [...] Chinese to English news task" (Hassan *et al.* 2018: 21). En el mismo sentido, Castilho *et al.* (2018: 11) afirman que la clásica distinción entre traductores profesionales y automáticos parece estar diluyéndose de forma gradual. Sin embargo, los propios investigadores de la empresa Microsoft que afirmaban haber alcanzado la paridad humana reconocían que la calidad de las traducciones depende de la combinación de lenguas, los dominios y la cantidad (y calidad) de los datos de entrenamiento (Hassan *et al.* 2018). En este sentido, Fomicheva *et al.* (2020) distinguen entre escenarios de altos y bajos recursos, en función de los pares de lenguas, y asocian un número relativamente escaso de datos de entrenamiento con traducciones de una

2 "At its core, NMT is based on the concept of deep learning, where a neural network with multiple layers processes the input data. For translation, this involves training the network on a large corpus of bilingual text so it can learn to predict the probability of a sequence of words in the target language, given a sequence of words in the source language" (https://deepai.org/machine-learning-glossary-and-terms/neural-machine-translation; consultado 28.6.2024).

3 De hecho, Gunathilaka y Ariyaratne (2019) comprueban que, incluso en una combinación *a priori* menor como inglés-cingalés (Sri Lanka), las traducciones de Google Translate con posedición generan traducciones precisas, si bien en textos literarios (prosa y verso) los traductores humanos muestran una mayor creatividad.

4 Way señalaba ya en 2018 que, en comparación con sus antecesores (Phrase-Based Statistical Machine Translation systems – PBSMT), los NMT habían reducido los errores morfológicos en un 19 %, los léxicos en un 17 % y los relacionados con el orden de palabras en un 50 %, algo que se evidenciaba en una reducción del número de correcciones necesarias durante el proceso de posedición de aproximadamente un 25 % (2018: 171).

calidad significativamente menor. Podemos ilustrar esta cuestión con los resultados obtenidos por Taira *et al.* (2021) a partir de la traducción generada por Google Translate de veinte instrucciones de alta hospitalaria frecuentemente empleadas en el servicio de urgencias. Veinte anotadores bilingües evaluaron, entre marzo de 2019 y febrero de 2020, traducciones automáticas del inglés a siete lenguas de uso común en los servicios de urgencias de Los Ángeles, California (EE. UU.). Los resultados evidencian que algunas combinaciones de lenguas obtienen traducciones relativamente precisas, concretamente los textos en español (94 % de las instrucciones evaluadas), tagalo (90 %), coreano (82,5 %), chino (81,7 %) y vietnamita (77,5 %). Sin embargo, otras lenguas mostraron datos preocupantes, en referencia al farsi y al armenio, con un porcentaje de precisión del 67,5 % y el 55 % respectivamente (Taira *et al.* 2021: 3362-3364). Incluso DeepL –sistema NMT que afirma ofrecer "la mejor traducción automática del mundo" y ser, en combinaciones como inglés-alemán e inglés-japonés, seis veces más preciso que sus competidores[5]– reconoce en su página web que su sistema tiene puntos débiles, especialmente en relación con la jerga especializada y la terminología de determinados sectores.[6] Otro dominio que se considera problemático es la literatura, que, en opinión de Wang *et al.* (2023), supone el mayor reto para la traducción automática a causa de su compleja naturaleza. Otro factor que tiene un impacto considerable en la calidad de las traducciones automáticas serían las diferencias estructurales entre las lenguas (Manakhimova *et al*. 2023).

3. Evaluación de la calidad de la traducción automática

No hay espacio en este breve artículo para presentar en detalle la cuestión de la evaluación de la calidad de la traducción (*translation quality assessment* – TQA).[7] Definir el mismo concepto de *calidad de la traducción*, así como establecer el modo adecuado de medirla, supone todo un reto a causa precisamente de la complejidad del propio proceso traductológico (Castilho *et al.,* 2018: 10). Sin embargo, se trata de un elemento muy relevante tanto desde la perspectiva teórica como pedagógica e incluso industrial.

Respecto a las traducciones generadas por NMT, a los métodos tradicionales de evaluación de la calidad, realizada principalmente por comparación del TO y el TM por parte de traductores profesionales, se les han ido sumando en los últimos años distintos programas que, como BLEU (Bilingual Evaluation Understudy), TER (Translation Error Rate) o METEOR (Metric for Evaluation of Translation with Explicit Ordering), evalúan automáticamente los resultados. En cuanto a la fiabilidad de estas tecnologías, Castilho *et al.* (2018: 26) se muestran escépticos, dado que: "one can claim that automatic MTE [MT Evaluation] metrics provide scores that appear to be objective and reliable, but the way in which they work is based on a number of assumptions that can raise some concerns as to their actual value". Un buen ejemplo de la problemática asociada a estos programas sería que la medición de la calidad de las traducciones automáticas se basa en la similitud de estas respecto a traducciones de referencia realizadas por humanos

5 *Cfr.* https://www.deepl.com/es/whydeepl (consultado 22.2.2024).

6 *Cfr.* https://www.deepl.com/es/blog (consultado 22.2.2024).

7 Puede leerse una presentación de esta problemática, a modo de ejemplo, en House (2001), Martínez Melis y Hurtado Albir (2001) o, más reciente, en Castilho *et al.* (2018).

profesionales. Sin embargo, resulta innegable que la traducción no es un proceso unívoco y que, para un mismo TO, puede haber distintos TM igualmente válidos.

En general, puede afirmarse que, pese a los intentos por parte de traductólogos e industria, en los procesos de TQA sigue existiendo una escasez de estandarización y un exceso de inconsistencia, y eso a pesar de las apelaciones de varios especialistas a la búsqueda de criterios objetivos para valorar las traducciones como los promulgados por Martínez Melis y Hurtado Albir (2001) hace ya más de dos décadas. Entre los escasos puntos en común en este debate teórico y metodológico se halla la división del proceso de calidad de las traducciones generadas por MT en dos elementos diferenciados: *precisión* y *fluidez*. La *precisión (accuracy),* conocida también como *adecuación (adequacy),* estaría orientada al TO y al proceso traductológico en sí y puede definirse, en general, como "the extent to which the translation transfers the meaning of the source-language unit into the target" (Castilho *et al.* 2018: 18). Por el contrario, la *fluidez (fluency)* está claramente orientada al TM, hasta el punto de que es independiente del TO, y puede definirse como: "the extent to which the translation follows the rules and norms of the target-language (regardless of the source or input text)" *(idem).* El análisis de la *fluidez* –tal como puede observarse, a modo de ejemplo, en el estudio de Specia y Shah (2018: 223-224)– juzga elementos como la ortografía, la tipografía, la gramática (concordancia, tiempo y modo verbales, orden de palabras, uso de artículos y preposiciones...) o la inteligibilidad del TM. Así, mientras juzgar la *precisión* exige tener un elevado nivel de competencia en las dos lenguas involucradas, para valorar la *fluidez* la competencia es necesaria únicamente en la lengua del TM (Castilho *et al.* 2018: 18). En cualquier caso, a pesar de ello, la gran mayoría de los estudios que juzgan la TQA analizan tanto la *precisión* como la *fluidez* de las traducciones mediante anotadores bilingües o traductores profesionales. Esta división entre precisión y fluidez, sin embargo, resulta fundamental para nuestra propuesta metodológica, ya que, como explicaremos detalladamente en la próxima sección de este artículo, nuestra aportación se centra en la fluidez y descarta por completo la evaluación de la precisión.

Otra aproximación teórica que nos ha servido de inspiración es el concepto de *traducción centrada en el usuario* (user-centered translation – UCT) propuesta por Suojanen, Koskinen y Tuominen con la intención de "emphasize the central role of the user, or reader, in the translation process" (Suojanen *et al.* 2014: 1). La UCT propone que el proceso de producción debe tener en cuenta al usuario o lector y debe ser, por tanto, de carácter interactivo. En consonancia con lo anterior, esta perspectiva, afín a las teorías traductológicas funcionales, le otorga gran importancia a la *usabilidad (usability),* es decir, al hecho de que el TM resulte efectivo, eficaz y satisfactorio para el usuario.[8] En cuanto a la valoración de la calidad, esta incluye factores que, como el contexto cultural, van mucho más allá de las posibilidades de nuestra propia propuesta metodológica de investigación (ver sección 4). Sin embargo, también aporta elementos de análisis que consideramos relevantes y, muy especialmente, en lo tocante al concepto de la ya citada

8 No se trata, en ningún caso, de un concepto exclusivo del campo de la TQA, como podemos observar en la definición que aporta ISO/TR 16982 de este concepto: "Usability (see ISO 9241-11) is the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. Effectiveness is fundamental as it is about achieving the intended goal(s). Efficiency is about the resources (such as time or effort) needed by users to achieve their goals so it can be important. In addition, it is important that users are satisfied with their experience, particularly where users have discretion over whether to use a product and can readily choose some alternative means of achieving their goals" (International Organization for Standardisation 2002).

usabilidad, de acuerdo con la cual, la lengua del TM debe ser "suitable for the purpose in terms of *style, register and terminology*" (Suokas 2019: 31; la cursiva es nuestra).

Algunos especialistas han investigado la usabilidad *aislada* del TM desde una perspectiva algo distinta, como los experimentos con una prueba realizada mediante *eye tracking* con traducciones de rompecabezas lógicos por parte de humanos y MT en la combinación inglés-danés (Klerke *et al.* 2015) o el estudio de Castilho y O'Brien (2016), igualmente basado en la tecnología de *eye tracking,* en el que se evalúa la usabilidad de textos en inglés y sus traducciones automáticas al alemán, tanto *en bruto* como tras un proceso de posedición.

4. Justificación de la propuesta teórica y metodología de análisis (evaluación basada en el usuario)

Si bien son numerosísimos los estudios publicados que evalúan la calidad de las traducciones automáticas, consideramos que aún sigue habiendo espacio para nuevas aproximaciones teóricas y metodológicas a esta problemática.

En primer lugar, porque la inmensa mayoría de estos trabajos analizan pares de lenguas con gran cantidad de datos o *high-resource language pairs* (Fomicheva *et al.* 2018) y, muy especialmente, traducciones del o al inglés en combinación con otra lengua *mayoritaria* como el francés, el alemán o el chino (Ranathunga 2023: 12). Por supuesto, también han salido a la luz estudios parcial o completamente dedicados a combinaciones entrenadas con un volumen de datos mucho menor, es decir, *low-resource language pairs,* como sería el caso, a modo de ejemplo, del inglés-farsi e inglés-armenio comentado en la sección 2 (Taira *et al.* 2021) o de las traducciones literarias en la combinación inglés-catalán (Toral y Way 2018). Sin embargo, existe una sorprendente (e incluso preocupante) escasez de trabajos de investigación sobre TQA en combinaciones no *anglocéntricas*. Entre las excepciones podemos nombrar los estudios que analizan las combinaciones marati-hindi (Ul Haq *et al.* 2020) o italiano-rumano (Lakew *et al.* 2018). Hendy *et al.* (2023), por su parte, analizan traducciones en dieciocho pares con y sin inglés como lengua de referencia, si bien su estudio se centra en los sistemas *Generative Pre-trained Transformer*, como chatGPT, y no en las aplicaciones diseñadas específicamente para tareas de traducción.

Por otra parte, como ya hemos indicado, la mayoría de los análisis ponen el foco en la evaluación del proceso de traducción mediante la comparación del TO y el TM, generalmente mediante valoraciones realizadas por traductores profesionales, lo que resulta razonable, ya que estos profesionales han sido formados para la labor traductora y, además, cuentan con experiencia en el campo objeto de análisis. Así, tal y como señalan Castilho *et al.* (2018: 23): "professionals can be assumed to provide more reliable results". Sin embargo, creemos que las investigaciones de este perfil metodológico pueden (y deben) ser complementadas con una perspectiva de análisis distinta que se centre en la percepción de la calidad del TM por parte del usuario no profesional, es decir, de sus potenciales lectores o, como los denominan Castilho y O'Brien (2016), sus destinatarios finales (*end users*). Estos, como reconocen Castilho *et al.* (2018: 23), pueden resultar igual de útiles en algunas tareas de TQA que los traductores profesionales. Así, y en consonancia con los conceptos de fluidez y usabilidad (ver sección 3), consideramos importante que la calidad de la traducción sea también evaluada como texto independiente, dado que el lector,

por ejemplo, de una página web no siempre es consciente de que está ante un texto traducido y, mucho menos aún, de si se trata de una traducción realizada por un humano o por un sistema NMT. Consideramos que la experiencia "aislada" del usuario también resulta altamente relevante y que esta debe ser igualmente tenida en cuenta a la hora de abordar la problemática de la calidad de las traducciones. En este sentido, los traductores evalúan de un modo sistemáticamente distinto a como lo hace un anotador no profesional, el verdadero público objetivo del producto. Por supuesto, siguiendo esta premisa, también las condiciones del experimento y los elementos a evaluar deben ser distintos, como detallaremos un poco más adelante. Se trata, en cualquier caso, de complementar –y, en ningún caso, reiteramos, sustituir– la evaluación *tradicional* de los textos producidos por NMT.

Partimos de la premisa de que –salvo en condiciones más bien excepcionales como las del experimento de Hassan *et al.* (2018) sobre la traducción de noticias del chino al inglés o en aquellas traducciones automáticas que han sido objeto de un proceso eficaz de posedición (Castilho y O'Brien 2016; Gunathilaka y Ariyaratne 2019)– los textos generados por los traductores automáticos presentan una calidad marcadamente inferior que los escritos por nativos o los traducidos por profesionales, especialmente en combinaciones no anglocéntricas. Por eso, consideramos necesario investigar si el usuario/lector percibe estas traducciones como textos de baja calidad o, por el contrario, como textos que presentan una calidad *suficiente*. En este sentido, Way (2018: 159) da por hecho que los usuarios de NMT están satisfechos con estas traducciones –por mucho que, como también señala, su calidad esté lejos de ser la ideal– porque, de lo contrario, su uso no estaría generalizado a escala global. Creemos que esta perspectiva de investigación ha sido insuficientemente explorada hasta el momento y que, al mismo tiempo, resulta muy relevante porque, en la actualidad, estamos expuestos a traducciones generadas por NMT de forma habitual, en muchos casos sin saber (e incluso sin importarnos en nuestro papel de usuarios) si se trata de textos originales o traducidos. De hecho, es probable que una buena parte de la sociedad perciba estas traducciones como un elemento más del corpus de lectura de su propia lengua. Además, la generalización en nuestro día a día de esos textos de calidad inferior (no necesariamente de calidad ínfima, pero sí menor que la de los textos originales y de las traducciones profesionales) podría tener un impacto significativo en la sociedad. Y no se trata solo de los potenciales malentendidos, confusiones e incluso distorsiones de la información, sino que, si los hablantes se acostumbran a leer textos poco fluidos o mal estructurados, no podemos descartar que acaben adoptando esas características en su propia comunicación escrita.

Una vez aclarados estos aspectos sobre la justificación de nuestra propuesta teórica de evaluación de la calidad de las traducciones generadas por sistemas NMT, pasamos a presentar en detalle la metodología de nuestro análisis de carácter aún exploratorio:

- Partimos de una combinación de lenguas no anglocéntrica con, *a priori,* una cantidad relativamente modesta de material de entrenamiento. Lo hacemos así, primero, porque estas combinaciones han recibido una atención mucho menor que aquellas en que se incluye el inglés y, segundo, porque los datos obtenidos para estos pares de lenguas han demostrado que las traducciones presentan una menor calidad que las relativas a *high-resource language pairs*. Concretamente, analizaremos traducciones del checo al español, entre otras razones, porque somos traductores de esta combinación y conocemos bien el mercado. Así, contamos

con información fidedigna de que las tres traducciones "humanas" que han sido objeto de análisis fueron realizadas por traductores profesionales sin ayuda de sistemas NMT.

- Dado que distintos especialistas proclaman que la calidad de los resultados producidos por los sistemas NMT depende no solo de la combinación de lenguas, sino también del género o tipología del TO, en nuestro análisis los anotadores evaluaron tres tipos de textos: un texto técnico (las instrucciones de uso y la composición de un producto de belleza), uno de *marketing* (un blog publicado en una web de productos de belleza con motivo del Día de San Valentín) y uno literario (el comienzo de la versión española de la novela de Irena Dousková (2023) *El baile del Oso*). En todos los casos, los fragmentos checos cuya traducción fue evaluada tienen una extensión de aproximadamente cien palabras y fueron presentados a los anotadores en orden aleatorio. Los dos primeros textos se tradujeron completos (con la intención de que los sistemas NMT tuvieran a su disposición la mayor información posible sobre el contexto), si bien luego se escogió tan solo un fragmento para que la extensión de los tres textos fuera similar. En el caso de la novela, dado que no era posible traducirla en su integridad, se escogió, como ya hemos señalado, su comienzo: la parte de la obra que, *a priori*, menos contexto necesita para su correcta traducción.
- De modo similar a otros muchos autores, en nuestro experimento combinamos traducciones humanas y de sistemas NMT. Concretamente, las propuestas por DeepL y Google Translate (los mejores NMT de acuerdo con la investigación realizada por un grupo independiente de investigadores para DeepL en 2020 y 2021).[9] En cuanto a la tercera traducción automática, elegimos el *chatbot* ChatGPT desarrollado por OpenAI. Más concretamente, su versión gratuita en el momento de realizar las traducciones para el experimento: ChatGPT 3.5. Si bien no se trata de un sistema NMT *per se* y no está desarrollado específicamente como aplicación de traducción, en este artículo y con el objetivo de simplificar la redacción, ChatGPT será también denominado sistema NMT, dado que "[it] employs a sequence-to-sequence model, a type of neural network architecture commonly used in machine translation tasks" (Dalayli 2023: 22). Resulta importante señalar que su empleo como herramienta de traducción se ha extendido con rapidez, especialmente entre los jóvenes (Sahari *et al.* 2023) e incluso han sido publicadas ya varias investigaciones que analizan la TQA de los textos que genera. En este sentido, Lee (2023) afirma que la calidad de las traducciones de ChatGPT, a pesar de no haber sido diseñado específicamente para este propósito, igualan o incluso superan a las generadas por las aplicaciones de traducción automática. Sin embargo, Jiao *et al.* reducen considerablemente el alcance de esta afirmación: "[...] ChatGPT performs competitively with commercial translation products (e.g., Google Translate) on high-resource European languages but lags behind significantly on low-resource or distant languages" (2023: 8). Hendy *et al.* (2023), por su parte, destacan las ventajas del uso de modelos GPT (Generative Pre-trained Transformer) para tareas de traducción, ya que, gracias a su capacidad de aprendizaje, resulta sencillo mejorar la calidad de los textos resultantes. Además, añaden que, a diferencia de los sistemas NMT, es posible modular el estilo o ciertos matices de las traducciones generadas mediante el diseño eficaz de *prompts*, es decir, proporcionándole al modelo de lenguaje indicaciones específicas para guiar su respuesta, lo que, de acuerdo con Gao *et al.* (2023), puede "fully

9 *Cfr.* https://www.deepl.com/es/whydeepl (consultado 22.2.2024).

unleash the translation power of ChatGPT". En nuestro caso, para la tarea de traducción se incluyeron dos informaciones en un único *prompt:* la dirección de la traducción (checo-español) y el contexto o tipología textual de la traducción (texto literario, de *marketing* o técnico). En cuanto a DeepL, empleamos la versión de pago (DeepL pro), que permite escoger, a diferencia de su versión gratuita y de Google Translate, entre registro "formal" e "informal". En el caso del texto técnico y del fragmento de la novela, escogimos "formal"; para la traducción del texto de *marketing,* seleccionamos "informal", de acuerdo con las propias indicaciones del cliente para las traducciones de sus blogs al español. Por otra parte, nuestra investigación aporta una novedad metodológica innovadora en el campo de la TQA: mientras que la mitad de los anotadores (Grupo A) valoró los cuatro textos (tres producidos por NMT y uno por humanos), la otra mitad (Grupo B) evaluó las traducciones de NMT sin contar con el modelo de la traducción humana profesional. De este modo, pretendemos sacar a la luz información sobre la medida en que los usuarios de los textos producidos por NMT perciben la calidad de estas traducciones de forma aislada –del mismo modo que las leerían en una web, por ejemplo– sin una traducción supuestamente de elevada calidad con la que compararla.

- Los anotadores tuvieron acceso a las tres o cuatro traducciones (dependiendo de si formaban parte del Grupo A o B) de cada texto original checo al mismo tiempo, de modo que podían evaluarlas a través de un proceso comparativo. Por otra parte, y a diferencia de algunas investigaciones que emplean un sistema de *ranking* (Toral y Way (2018), DeepL[10]), a los anotadores se les proporcionó una escala de valoración Likert, mediante la que se expresa el nivel de acuerdo (o desacuerdo) en una escala simétrica para una serie de afirmaciones. La escala Likert, que valora dichas afirmaciones en una escala de 1 (no estoy nada de acuerdo con la afirmación) a 5 (estoy muy de acuerdo con la afirmación), es una medida frecuentemente empleada en los estudios cuantitativos de TQA para evaluar la adecuación y la fluidez (Castilho *et al.* 2018: 18). La escala se aplicó para las siguientes variables: a) *fluidez*, es decir, la naturalidad con que se lee la traducción en la lengua meta; b) *corrección gramatical*, incluida la coherencia del TM, y que, en la literatura especializada, suele considerarse un subelemento de la fluidez (ver sección 3); y c) *usabilidad,* que, partiendo de los principios propuestos por Suojanen, Koskinen y Tuominen (2014), hace referencia principalmente a la adecuación del vocabulario y el estilo de la traducción a su intención comunicativa.[11] Cada anotador aportó un juicio (o ítem) por cada una de las tres variables (*fluidez, corrección gramatical* y *usabilidad*) y para cada una de las (tres o cuatro) versiones de los tres textos traducidos del checo al español. Así, los diez miembros del Grupo A anotaron nueve textos y los diez del Grupo B evaluaron doce textos, generando un total de 630 juicios.
- El experimento se realizó en febrero de 2024 a través de formularios Google Forms. Participaron veinte hablantes nativos de español mayores de 18 años que, en su inmensa mayoría,

10 *Cfr.* https://www.deepl.com/es/whydeepl (consultado 22.2.2024).

11 Dado que los anotadores desconocen la terminología de la teoría traductológica, las afirmaciones se formularon de un modo menos técnico: a) El texto se lee con fluidez, no hay expresiones extrañas o poco naturales en español; b) El texto está libre de errores gramaticales e incoherencias; c) En tu opinión como hablante nativo, el vocabulario y el estilo son los adecuados para la intención comunicativa de un... c1) texto de *marketing*: que el lector compre productos de belleza; c2) texto técnico: informar al lector de la composición y las instrucciones de uso del producto; c3) texto literario.

contaban con estudios superiores y, aunque la muestra fue heterogénea, más de la mitad de ellos eran estudiantes de doctorado o profesores universitarios en estudios de humanidades. Además, diez participantes eran hablantes de español peninsular y otros diez de distintas regiones de Hispanoamérica: tres de México, tres de Colombia, dos de Argentina y dos de Uruguay (también se preguntó por el género de los anotadores, formado por diez hombres y diez mujeres, si bien para este análisis piloto no lo consideramos una variable de investigación). Esta decisión metodológica se tomó para comprobar si había diferencias entre el grado de satisfacción de hablantes nativos europeos y americanos respecto a las traducciones realizadas por humanos y por sistemas NMT. Hay que tener en cuenta que los traductores profesionales realizaron sus encargos para un público objetivo de español europeo: una editorial barcelonesa y la versión española (con dominio .es) de una web checa de productos de belleza. Por otra parte, ninguno de los sistemas NMT empleados permite escoger entre diversas variedades del español (DeepL solo distingue dos variedades diatópicas: inglés americano y británico) y a ChatGPT no se le hizo mención de la variedad que debía emplear para las traducciones. Sin embargo, una simple prueba con una oración inglesa que contenga las palabras *cars* y *peaches* muestra que DeepL genera traducciones propias del español peninsular (*coches* y *melocotones*) de forma consistente, mientras que Google Translate varía según la formulación de la oración: en algunas ocasiones lo traduce como *coches* y *melocotones* (I like cars and peaches) y, en otras, con las variedades americanas *auto* y *durazno* (I will never forget the first car I bought and the first peach I ate). Por su parte, ChatGPT 3.5 parece generar siempre términos americanos: *auto/carro* y *durazno.* Al ser cuestionado por el uso de esta variedad americana, ChatGPT responde que lo ha hecho por una cuestión de *consistencia* y *accesibilidad*. Creemos que la razón para estas diferencias puede hallarse en el hecho de que tanto Google Translate como ChatGPT son entrenados con enormes cantidades de datos sin un nivel exigente de cribado y, por tanto, las variedades americanas tienen mayor peso o incluso se producen variaciones (el caso de Google Translate). Por el contrario, DeepL ha desarrollado un sistema de redes neuronales que se basa en una menor cantidad de datos, pero escogidos de un modo más selectivo, probablemente dando prioridad al español europeo.[12] Lo que también ha quedado demostrado, a través de la repetición de las traducciones empleadas en el experimento en ordenadores con IP americana (en Colombia y Argentina, concretamente), es que las traducciones generadas por DeepL y Google Translate no varían dependiendo de la región del mundo desde donde se soliciten.

- El análisis tiene un carácter preeminentemente cuantitativo y aún experimental, con una cantidad de datos relativamente pequeña, por lo que se buscan tendencias y no resultados estadísticamente significativos. Además, se entrevistó a la mitad de los anotadores con la finalidad de recopilar información cualitativa que ayudara a interpretar los resultados y a formular con mayor precisión los futuros experimentos pensados para ser anotados por cientos de participantes.

Antes de pasar a exponer los resultados del experimento, enumeramos las hipótesis de investigación basadas en los datos apuntados hasta ahora en este artículo.

12 *Cfr.* https://www.deepl.com/es/blog/how-does-deepl-work (consultado 22.2.2024).

- Hipótesis 1: Los anotadores adjudicarán a las traducciones humanas valores más elevados en todos los tipos de texto y en las tres variables (fluidez, corrección gramatical y usabilidad). El sistema automático mejor valorado será DeepL, en consonancia con los datos publicados en su web, seguido por Google Translate. El peor valorado será ChatGPT, dado que no es un sistema diseñado específicamente para este fin. Sin embargo, consideramos posible que los resultados dependan de la tipología textual y/o de las variables.
- Hipótesis 2: Los hablantes nativos europeos evaluarán con una puntuación más elevada las traducciones humanas y de DeepL por estar enfocadas al usuario europeo. Por el contrario, los anotadores hispanoamericanos puntuarán mejor las traducciones generadas por ChatGPT. La inconsistencia documentada en la variedad de español para los resultados propuestos por Google Translate hace que no podamos formular una hipótesis a este respecto.
- Hipótesis 3: Los diez anotadores que han leído las cuatro traducciones (incluida una realizada por un traductor profesional) asociarán valores más bajos a las traducciones generadas por los sistemas NMT, dado que contarán con un referente de supuesta mayor calidad. Por el contrario, los participantes que solo valoren las traducciones automáticas les asociarán, de modo general, una mayor calidad.
- Hipótesis 4: Las traducciones a cargo de tecnologías NMT del texto técnico recibirán una puntuación más elevada que las del texto de *marketing,* que emplea técnicas de persuasión y uso de lenguaje figurado, y estas, a su vez, obtendrán una mejor valoración que las del fragmento literario por tratarse del mayor reto para la traducción automática.

5. Presentación e interpretación de los resultados

En las Tablas 1–3 (ver pp. 79 y 80) podemos observar los resultados obtenidos de la evaluación proporcionada por los diez anotadores del Grupo A (cinco europeos y cinco americanos), es decir, aquellos que leyeron las tres traducciones generadas por sistemas NMT y la traducción humana de referencia. Cada una de las tablas expone los datos de uno de los tres textos traducidos de acuerdo con las tres variables de evaluación analizadas. También se presenta el valor medio de la suma de las calificaciones de las tres variables.

En cuanto al texto técnico (las instrucciones de uso y la composición de un producto de belleza), los datos señalan que, en las tres variables (fluidez, corrección y usabilidad), la traducción mejor valorada es siempre la humana, destacando muy especialmente en *fluidez*, para la que obtiene una nota media que roza la *perfección:* 4,8. Además, la aplicación DeepL muestra valoraciones muy elevadas y consistentes (siempre iguales o superiores al 4) en las tres variables y, por tanto, se muestra como un sistema NMT muy fiable para traducciones de esta tipología textual en la combinación checo-español. Las evaluaciones de Google Translate, por su parte, lo estiman como una aplicación que proporciona resultados de alta calidad en términos de *fluidez* (4,1) y *usabilidad* (4,1), pero no así en *corrección gramatical* (3,3). De hecho, la traducción generada por ChatGPT supera a Google Translate tanto en términos de *usabilidad* (4,2) como, muy especialmente, de *corrección* (3,7). De acuerdo con los datos obtenidos, ChatGPT sería un traductor de

una calidad ligeramente superior a Google Translate (3,9 frente a 3,8 de media). En cualquier caso, las tres aplicaciones presentan unos valores relativamente similares entre sí, con medias en torno al 4.

Las valoraciones recibidas durante el experimento para el texto de *marketing* (un blog publicado en una web de productos de belleza con motivo del Día de San Valentín) muestran valores medios más bajos respecto al texto técnico en las cuatro traducciones (incluida también, por tanto, la humana), si bien esta disminución en la valoración resulta especialmente marcada en las aplicaciones DeepL y ChatGPT. Algunos de los patrones observados en la Tabla 1 se repiten: nuevamente la traducción mejor valorada es la humana para todas las variables y Google Translate muestra números elevados en *fluidez* (3,9) y *usabilidad* (4), pero no así en *corrección gramatical* (3,1). Es precisamente esta variable en la que los tres sistemas NMT muestran sus peores puntuaciones, muy por debajo de las obtenidas para el texto técnico. De hecho, la peor valoración de entre las recogidas en este estudio se refiere a la *corrección gramatical* del texto generado por ChatGPT, que apenas alcanza el 2,5. De acuerdo con estos datos, DeepL sería una aplicación fiable a la hora de generar traducciones de textos técnicos del checo al español, pero no tanto al trabajar con textos de *marketing*.

La Tabla 3 muestra los datos obtenidos para las cuatro traducciones de un texto literario (el comienzo de la versión española de la novela de Irena Dousková (2023) *El baile del Oso*). Nuevamente es la versión humana la mejor valorada. Sin embargo, también estamos ante la única variable para la que los anotadores puntuaron mejor un sistema NMT que una traducción profesional. Concretamente, nos referimos a la *fluidez* del texto creado por Google Translate, que obtuvo un 4,1 frente al 4 de la traducción humana y de la generada por DeepL. Llama la atención la relativamente elevada *corrección gramatical* (3,8) relacionada con Google Translate, que había mostrado cifras relativamente bajas para las dos tipologías textuales anteriormente comentadas (3,3 y 3,1 respectivamente). En el cómputo general, sin embargo, sobresale la traducción humana, muy especialmente gracias a la *usabilidad* (4,4). Por último, debemos señalar que Google Translate ha obtenido mejores puntuaciones que DeepL para las tres variables y que la aplicación ChatGPT, a pesar de haber sido informada en el *prompt* de que se trataba de una traducción literaria, generó un texto en español de una calidad sustancialmente inferior a la de los sistemas NMT.

En cualquier caso, y a pesar de la mayor calidad generalizada de la traducción humana, llama poderosamente la atención que las traducciones generadas por sistemas NMT hayan obtenido puntuaciones relativamente elevadas para las tres tipologías textuales. Así, si hiciéramos la media de todas las valoraciones obtenidas para cada uno de los tres textos, las traducciones humanas habrían obtenido un 4,3; DeepL y Google Translate, un 3,8; y ChatGPT, un 3,4: todas por encima de la calidad mínima aceptable, que en la escala Likert se situaría en un 3. De hecho, tan solo hay una variable con una puntuación que no alcanza el aprobado: la *corrección* del texto de *marketing* generado por ChatGPT (2,5).

De las cifras presentadas en la Tabla 4 (ver p. 80) destaca que todas las traducciones (incluida la humana) presentan sus peores resultados en la variable *corrección gramatical*. También es digno de mención que, tanto para *fluidez* como para *usabilidad*, las traducciones generadas por Google Translate superan a las de DeepL. Por último, hay que señalar que la variable de evaluación en la que más destaca ChatGPT es la *usabilidad*, con cifras cercanas a las obtenidas por los dos sistemas de traducción automática.

Los datos relativos a las variedades del español (Tabla 5; ver p. 81) muestran que los hablantes procedentes de países americanos del grupo A (concretamente dos mexicanos, dos uruguayos y un colombiano) han evaluado todas las traducciones generadas por sistemas NMT con puntaciones más elevadas que los nativos de la variedad europea de español. En algunos casos, estas divergencias son muy marcadas, especialmente en relación con el texto literario y el de *marketing,* documentándose entre las dos variedades de español cifras más cercanas en lo referente al texto técnico. El caso más extremo es el de la traducción del texto literario generado por ChatGPT, que obtuvo una puntuación media de 4,1 entre los hablantes americanos y de apenas 2,6 entre los europeos. Respecto a las traducciones humanas, la situación es la contraria: los anotadores europeos les han otorgado mayores puntuaciones que los americanos. Sin embargo, estas diferencias son mucho menos extremas que en el caso de las generadas por sistemas NMT y, de hecho, la más relevante se refiere al texto de *marketing,* anotado con una valoración de 4,5 por los hablantes europeos y de 4,1 por los americanos. Si tuviéramos en cuenta todos los textos, variables y tipos de traductor, los anotadores europeos puntuaron las traducciones con un 3,6 de media, mientras que los americanos alcanzaron la cifra de 4,15. Un dato especialmente interesante es que la media de puntuación para los dos traductores automáticos coincide en las valoraciones de los anotadores de ambos lados del Atlántico: los hablantes europeos les otorgaron tanto a DeepL como a Google Translate una media de 3,4 frente a los 4,2 puntos obtenidos también para los dos sistemas NMT por los hablantes americanos.

Estos datos parecerían apuntar al uso de una variedad americana en las traducciones generadas por las tres aplicaciones objeto de estudio. Sin embargo, esta no puede ser la causa, puesto que DeepL realiza traducciones a la variedad europea, tal y como demostrarían las búsquedas que dan como resultado *auto/carro/coche* y *durazno/melocotón* presentadas en la sección anterior. Así, consideramos más probable que los resultados se deban a que los anotadores americanos tienen una mayor tolerancia respecto a la calidad de los textos. En este sentido, no podemos descartar que pueda jugar un papel importante el hecho de que sean consumidores, en mucha mayor medida que los hablantes europeos, de lo que se ha denominado *español (neutro) latino* (López González 2019). Esta tolerancia puede observarse, además, en que valoran las traducciones humanas –indudablemente enfocadas a los hablantes de español europeo– con prácticamente las mismas clasificaciones que los anotadores procedentes de España. Al ser preguntados por este asunto, los anotadores americanos consultados reconocieron que habían considerado que algunas de las *anomalías* probablemente se debieran a que se trataba de textos escritos en español europeo. Esto, en nuestra opinión, podría haber provocado que valoraran traducciones automáticas de baja calidad con una puntuación relativamente elevada. En el futuro, las afirmaciones de una parte de los anotadores especificarán que deben valorar las variables respecto a su variedad de español o, al menos, respecto al español latino (al que sí están acostumbrados, como hemos señalado arriba), lo que podrá sacar a la luz su grado de satisfacción con los textos como usuarios de una región concreta. En cualquier caso, no debemos perder de vista que contamos aún con datos muy escasos: los aportados tan solo por cinco anotadores para cada variedad del español. De hecho, los resultados obtenidos de los diez anotadores del Grupo B, que leyeron solo las traducciones de tres textos, muestran la tendencia contraria: los hablantes europeos

otorgaron puntuaciones ligeramente más elevadas (3,4 de media total) que los americanos (3,2).[13] A todo esto hay que recordar el hecho innegable de que no existe una única variedad americana del español (Černý 2014: 25), lo que complica la obtención e interpretación de los datos.

La última cuestión que pretendía aclarar este estudio piloto se refería a si la presencia de una traducción de referencia (humana) de calidad supuestamente elevada[14] podía influir las valoraciones de los usuarios/lectores de traducciones generadas por NMT sin posedición. Por ello, la mitad de los anotadores (Grupo A) tuvo acceso a cuatro traducciones, incluida una humana, mientras que la otra mitad (Grupo B) evaluó tan solo aquellas generadas por sistemas NMT. La Tabla 6 (ver p. 81) muestra los resultados de la comparación entre las puntuaciones de los textos aportadas por los Grupos A y B. En contra de lo esperado, las evaluaciones del Grupo B fueron sistemáticamente más bajas que las del Grupo A. La única excepción es la valoración de la *corrección gramatical* de la traducción generada por DeepL, que fue ligeramente mejor puntuada por el Grupo B que por el Grupo A. En general, las divergencias entre las cifras obtenidas para los dos grupos son relativamente pequeñas, con la excepción de ChatGPT, aplicación esta para la que se evidencian diferencias más importantes. Debemos reconocer que no tenemos una explicación convincente para estos resultados, ya que suponíamos que los anotadores del Grupo A se verían influidos por las traducciones humanas de referencia y, por tanto, puntuarían las generadas por sistemas NMT con unos valores más bajos que los del Grupo B. Futuras investigaciones habrán de aclarar este aspecto.

13 En este caso, se trataba de dos hablantes argentinos, dos colombianos y uno mexicano.

14 Recordemos que la existencia de la supuesta "gold standard quality" de las traducciones humanas es un hecho, cuando menos, polémico (Castilho *et al.* 2018: 26).

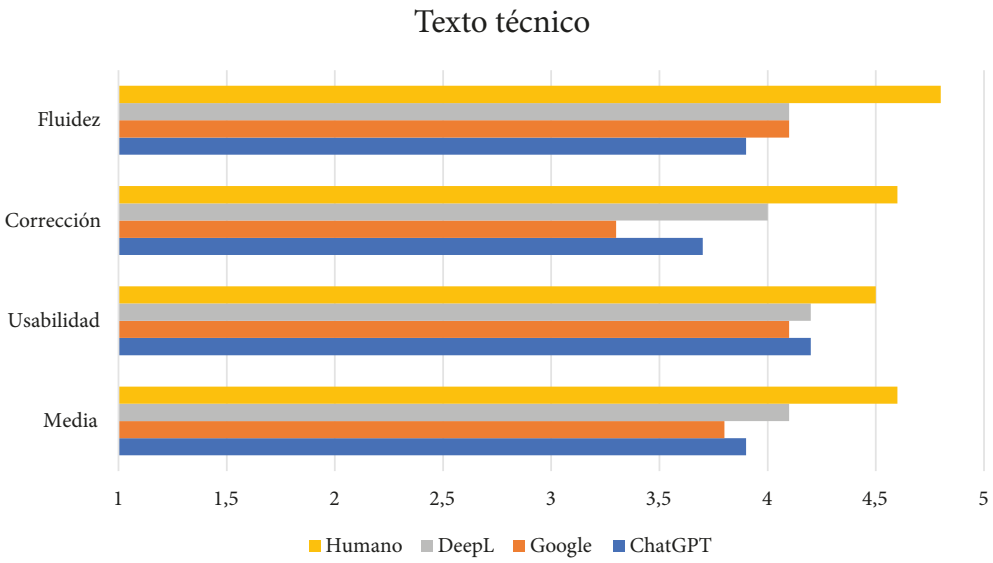

Tabla 1. Valoración obtenida para las traducciones del texto técnico

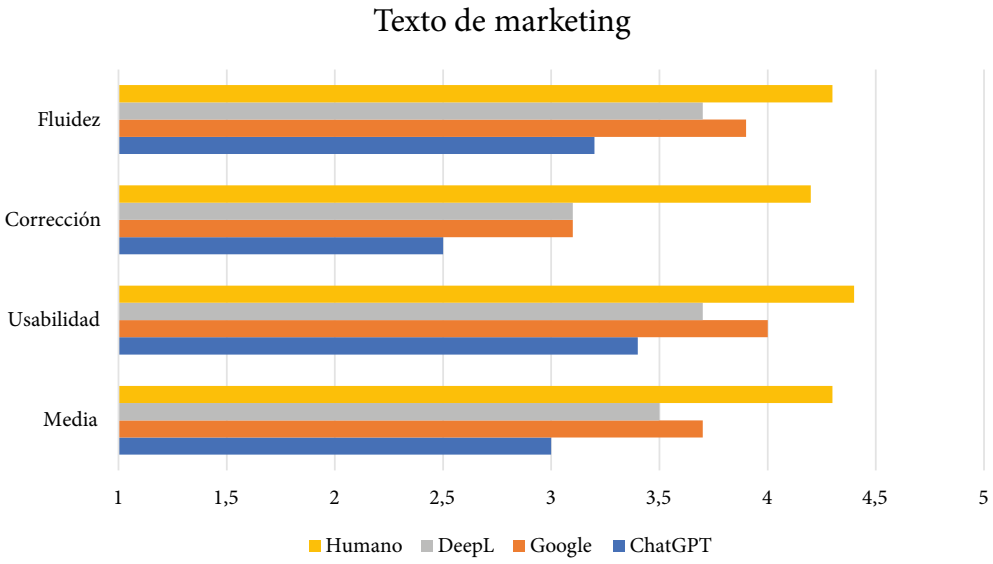

Tabla 2. Valoración obtenida para las traducciones del texto de *marketing*

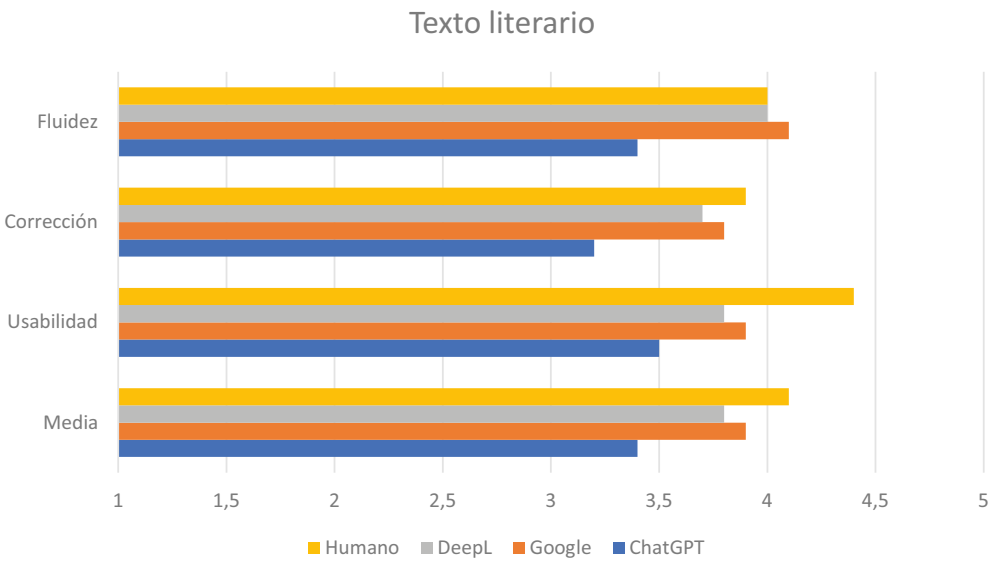

Tabla 3. Valoración obtenida para las traducciones del texto literario

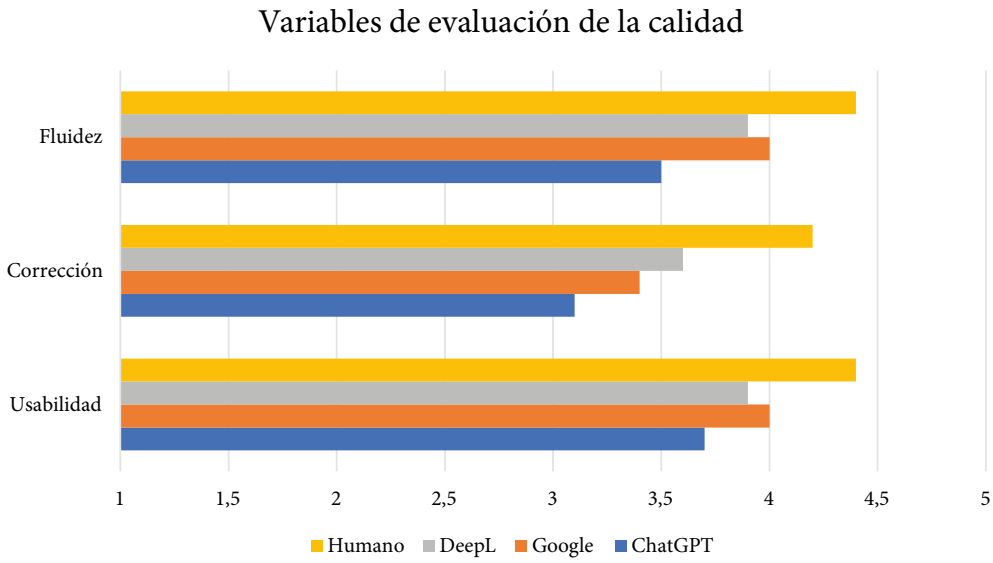

Tabla 4. Valoración de las variables de evaluación sin diferenciar la tipología textual

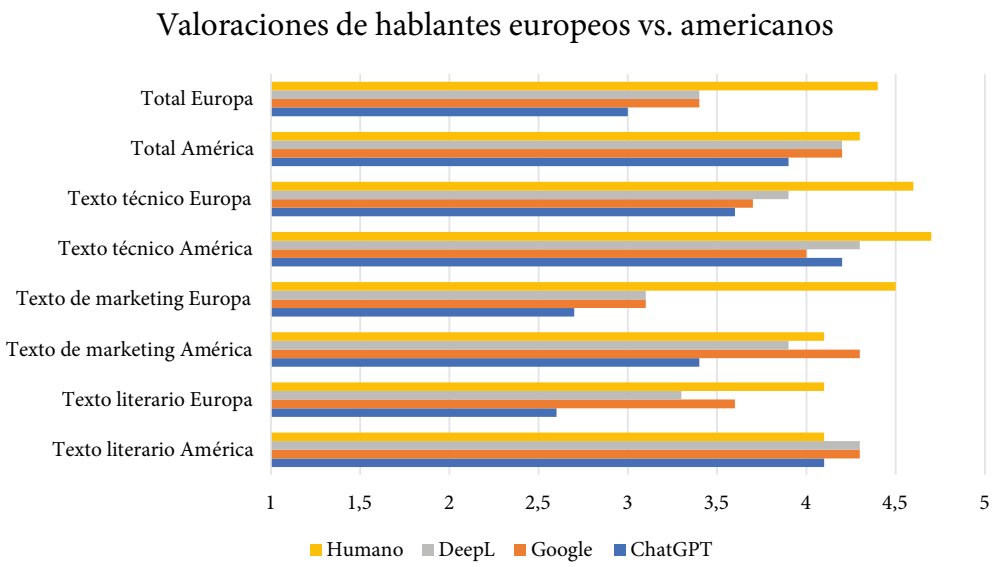

Tabla 5. Valoración de los anotadores de acuerdo con su variedad de español

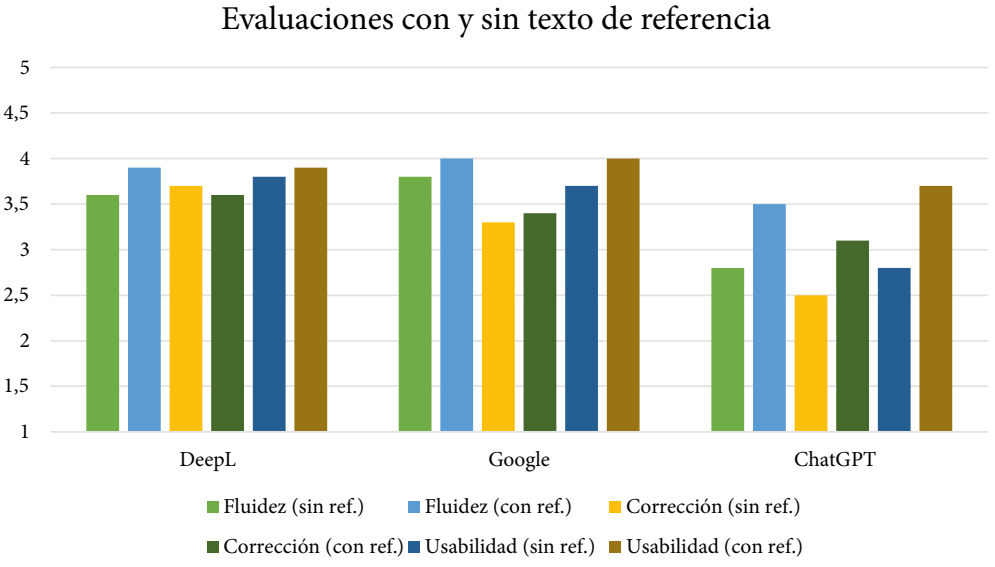

Tabla 6. Evaluación de las traducciones con y sin traducción humana de referencia

6. Conclusiones y discusión

La finalidad de este artículo, basado en un experimento piloto de características aún modestas, se limita a presentar una nueva propuesta teórica-metodológica y a mostrar algunas tendencias respecto a la calidad de las traducciones realizadas por humanos en contraste con las generadas por sistemas NMT en una combinación no anglocéntrica.

Si bien los datos obtenidos nos permiten responder con cierto grado de seguridad a algunas de las hipótesis planteadas al final de la sección 4, en general, nos generan muchas más dudas que certezas. Futuros experimentos más complejos, basados en un número de anotadores sustancialmente mayor y que tengan en cuenta nuevas variables (especialmente respecto a las variedades de español de los anotadores) habrán de confirmar estadísticamente las hipótesis y dar respuesta a las dudas generadas.

La Hipótesis 1 ha sido parcialmente confirmada. Efectivamente, los anotadores adjudicaron a las traducciones humanas valores más elevados que a los sistemas NMT en todas las tipologías textuales y en las tres variables de evaluación. La única excepción se refiere a la *fluidez* en el texto literario traducido por Google Translate, que obtuvo una puntuación una décima superior a la humana. Sin embargo, DeepL fue el sistema automático mejor valorado únicamente en el caso del texto técnico. Para las otras dos tipologías textuales, y en contra de la Hipótesis 1, fue Google Translate el que obtuvo valores iguales o más elevados en todas las variables. Así, la información publicada en la web de DeepL, según cual se trata de "la mejor traducción automática del mundo", difícilmente pueda ser aplicable a todas las combinaciones de lenguas y a todas las tipologías textuales. Por último, ChatGPT sí ha sido la aplicación peor valorada por los anotadores. Sin embargo, para el texto técnico ha demostrado ser capaz de crear traducciones de calidad (entendida esta en términos de *fluidez, corrección* y *usabilidad*), obteniendo una valoración ligeramente superior a la recibida por un sistema diseñado específicamente para este fin como es Google Translate.

La Hipótesis 2 no se ha visto confirmada. No se han apreciado diferencias sustanciales entre los anotadores europeos y americanos respecto a la valoración de la aplicación DeepL, que, *a priori,* hace uso de la variedad europea, por una parte, y Google Translate y ChatGPT, es decir, los sistemas NMT que supuestamente emplean una variedad americana para sus traducciones (ya sea de forma consistente o no), por la otra. Además, si bien es cierto que los hablantes nativos europeos evaluaron con puntuaciones más elevadas que los americanos las traducciones humanas dirigidas a un público español, las diferencias entre ambos grupos de anotadores son mínimas. Esto, unido al hecho de que los hablantes procedentes de países americanos evaluaron todas las traducciones generadas por sistemas NMT con puntaciones más elevadas que los nativos de la variedad europea, nos hace pensar en la posibilidad de que los lectores americanos tengan una mayor tolerancia respecto a la calidad de los textos que consumen.

La Hipótesis 3 tampoco se ha visto confirmada. Los diez anotadores del grupo A, que han leído las cuatro traducciones (incluida una realizada por un traductor profesional), asociaron valores más elevados a las traducciones generadas por sistemas NMT que los del grupo B, que no contaban con una traducción de referencia de (supuesta) elevada calidad.

La Hipótesis 4 ha sido parcialmente confirmada. Efectivamente, el texto con puntuaciones más elevadas ha sido el técnico. Sin embargo, la valoración de los anotadores de la traducción del

texto literario ha superado a la del texto de *marketing,* que ha generado valores muy bajos en la variable *corrección*.

Además de estas hipótesis, la investigación pretendía aclarar el grado de satisfacción de los hablantes nativos respecto a las traducciones generadas por sistemas NMT sin posedición. El resultado puede clasificarse de preocupante, ya que, si calculáramos la media de todas las valoraciones, las traducciones de DeepL y Google Translate habrían obtenido un 3,8, situándose, por tanto, más cerca de las calificaciones obtenidas por las traducciones humanas profesionales (4,3) que de la calidad mínima aceptable, que se situaría en un 3. Peor valoradas han sido las traducciones generadas por ChatGPT, que, sin embargo, superarían la barrera del aprobado con un 3,4 de media. Esto se halla en consonancia con la apreciación de Way (2018: 159), quien señala que los usuarios de MT probablemente estén satisfechos con estas traducciones porque, de lo contrario, no utilizarían estas aplicaciones. Hemos afirmado que estos datos resultan preocupantes. En este sentido, resulta necesario recordar que todos los especialistas, e incluso las propias empresas que desarrollan estas aplicaciones, coinciden en considerar que las traducciones generadas por la tecnología NMT sin posedición están, en general, aún lejos de igualar las creadas por traductores humanos profesionales, algo especialmente notable en combinaciones no anglocéntricas como checo-español. Podemos ejemplificar esta falta de calidad con un fragmento de la traducción a cargo de Google Translate del texto de *marketing* que obtuvo una puntuación total de 3,7:

> Elija una fragancia sensual para su pareja que impresione gratamente sus sentidos. Un regalo ideal para San Valentín será un perfume con temática de amor, como Giorgio Armani Sì Passione. ¡O busca un perfume unisex que haga que ambos huelan bien! De esta manera podrás demostrarles a todos que tú y tu pareja pertenecen el uno al otro.

Si bien se trata de conclusiones provisionales obtenidas de un experimento relativamente modesto, consideramos que estos datos deberían hacernos reflexionar sobre nuestra capacidad como sociedad para aceptar los textos de escasa calidad a los que estamos expuestos y, consecuentemente, sobre la necesidad de desarrollar el espíritu crítico ente los usuarios/lectores.

Referencias bibliográficas

Castilho, S.; Doherty, S.; Gaspari, F.; & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty, (Eds.). *Translation Quality Assessment* (pp. 9–38). Cham: Springer.

Castilho S.; & O'Brien, S. (2016). Evaluating the impact of light post-editing on usability. In N. Calzolari *et al.* (Eds.). *Proceedings of the tenth international conference on language resources and evaluation. Portorož, 23–28 May.* (pp. 310–316).

Černý, J. (2014). *El español hablado en América*. Olomouc: Univerzita Palackého v Olomouci.

Dalayli, F. (2023). Use of NLP Techniques in Translation by ChatGPT: Case Study. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)* (pp. 19–25). Varna (Bulgaria): INCOMA Ltd., Shoumen, Bulgaria.

Dousková, I. (2023). *El baile del oso*. Barcelona: La Fuga Ediciones.

Fomicheva, M.; Sun, S.; Yankovskaya, L.; Blain, F.; Guzmán, F.; Fishel, M.; Aletras, N.; Chaudhary, V.; & Specia, L. (2020). Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics, 8*, 539–555. <https://doi.org/10.1162/tacl_a_00330>

Gaspari, F.; Almaghout, H.; & Doherty, S. (2015). A survey of machine translation competences: Insights for translation technology educators and practitioners. *Studies in Translatology,* 23, 3, 333–358. <http://dx.doi.org/10.1080/0907676X.2014.979842>

Gao, Y.; Wang, R.; & Hou, F. (2023). How to Design Translation Prompts for ChatGPT: An Empirical Study. arXiv:2304.02182v2 [cs.CL]. <https://doi.org/10.48550/arXiv.2304.02182>

Gunathilaka, D. D. I. M. B.; & Ariyaratne, W. M. (2019). A Study on the Accuracy of Human Translation Output and Post-Edited Google Translate Output as far as English and Sinhalese Language Pair is considered: With Special Reference to Selected Literary and Non-literary Documents. *International Journal of Research and Innovation in Social Science (IJRISS), Volume III, Issue VII,* 503–510.

Hassan, H. *et al.* (2018). Achieving Human Parity on Automatic Chinese to English News Translation. arXiv:1803.05567 [cs.CL]. <https://doi.org/10.48550/arxiv.1803.05567>

Hendy, A.; Abdelrehim, M.; Sharaf, A.; Raunak, V.; Gabr, M.; Matsushita, H.; Kim, Y. J.; Afify, M.; & Awadalla, H. H. (2023). How good are gpt models at machine translation? A comprehensive evaluation. arXiv:2302.09210v1. <https://doi.org/10.48550/arXiv.2302.09210>

House, J. (2001). How do we know when a translation is good? In E. Steiner, & C. Yallop (Eds.). *Exploring Translation and Multilingual Text Production: Beyond Content* (pp. 127–160). Berlin: De Gruyter.

International Organization for Standardisation (2002). ISO/TR 16982:2002 ergonomics of human-system interaction—usability methods supporting human centred design. International Organization for Standardisation, Geneva. <https://www.iso.org/obp/ui/#iso:std:iso:ts:20282:-2:ed-2:v1:en> [29/2/2024]

Jiao, W.; Wang, W.; Huang, J.; & Wang, X. (2023). Is ChatGPT a good translator? Yes With GPT-4 As The Engine. arXiv:2301.08745v4. <https://doi.org/10.48550/arXiv.2301.08745>

Klerke, S.; Castilho, S.; Barret, M.; & Søgaard, A. (2015). Reading metrics for estimating task efficiency with SMT output. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning* (pp. 6–13). Lisbon: Association for Computational Linguistics.

Lakew, S. M.; Federico, M.; Negri, M.; & Turchi, M. (2018). Multilingual Neural Machine Translation for Low-Resource Languages, *IJCoL* [Online] (pp. 11–25). <https://doi.org/10.4000/ijcol.531>

Lee, T. (2023). Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Applied Linguistics Review* (Ahead of Print). <https://doi.org/10.1515/applirev-2023–0122>

López González, A. M. (2019). Español neutro – español latino: Hacia una norma hispanoamericana en los medios de comunicación. *Roczniki Humanistyczne,* 67, 5, 7–27. <https://doi.org/10.18290/rh.2019.67.5–1>

Manakhimova, S. *et al.* (2023). Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT),* 224–245. <doi.org/10.18653/v1/2023.wmt-1.23>

Martínez Melis, N.; & Hurtado Albir, A. (2001). Assessment In Translation Studies: Research Needs. *Meta*, 46, 2, 272–287. <https://doi.org/10.7202/003624ar>

Ranathunga, S.; Lee, E. A.; Skenduli, M. P.; Shekhar, R.; Alam, M.; & Kaur, R. (2023). Neural Machine Translation for Low-resource Languages: A Survey. *ACM Computing Surveys,* 55, 11, Article 229. <https://doi.org/10.1145/3567592>

Sahari, Y.; Al-Kadi, A. M. T.; & Ali, J. K. M. (2023). Cross Sectional Study of ChatGPT in Translation: Magnitude of Use, Attitudes, and Uncertainties. *Journal of Psycholinguistic Research 52*, 2937–2954. <https://doi.org/10.1007/s10936–023–10031-y>

Specia, L.; & Shah, K. (2018). Machine Translation Quality Estimation: Applications and Future Perspectives. In J. Moorkens, Sh. Castilho, F. Gaspari, & S. Doherty, (Eds.). *Translation Quality Assessment* (pp. 201–235). Cham: Springer.

Suojanen, T.; Koskinen, K.; & Tuominen, T. (2014). *User-Centered Translation.* London: Routledge.

Suokas, J. (2019). User-centered Translation and Action Research Inquiry. Bringing UCT into the Field. *Kääntämisen ja tulkkauksen tutkimuksen symposiumin verkkojulkaisu / Electronic Journal of the KäTu Symposium on Translation and Interpreting Studies, Vol. 12,* 29–43.

Taira, B. R.; Kreger, V.; Orue, A.; & Diamond, L. C. (2021). A Pragmatic Assessment of Google Translate for Emergency Department Instructions. *Journal of General Internal Medicine*, *Volume 36*, 3361–3365.

Toral, A.; & Way, A. (2018). What Level of Quality Can Neural Machine Translation Attain on Literary Text? In J. Moorkens, Sh. Castilho, F. Gaspari, & S. Doherty, (Eds.). *Translation Quality Assessment* (pp. 263–287). Cham: Springer.

Ul Haq, S.; Rauf, S. A.; Shoukat, A.; & Saeed, A. (2020). Document Level NMT of Low-Resource Languages with Backtranslation. *Proceedings of the 5th Conference on Machine Translation (WMT)*, online (pp. 442–446).

Wang, L. *et al.* (2023). Findings of the WMT 2023 Shared Task on Discourse-Level Literary Translation: A Fresh Orb in the Cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation (WMT)* (pp. 55–67). <https://aclanthology.org/2023.wmt-1.3.pdf>

Way, A. (2018). Quality Expectations of Machine Translation. In J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty, (Eds.). *Translation Quality Assessment* (pp. 159–178). Cham: Springer.

Zaretskaya, A.; Corpas Pastor, G.; & Seghiri, M. (2015). Translators' Requirements for Translation Technologies: a User Survey. In *New Horizons in Translation and Interpreting Studies* (pp. 247–254). Geneva: Tradulex.