Entreculturas 17 (2026) pp. 6-41— ISSN: 1989-5097

Traducción automática en contextos clínicos: los cuestionarios de autodiagnóstico

Machine translation in clinical contexts: self-assessment questionnaires

Malena Ruiz García-Casarrubios
https://orcid.org/0009-0002-1508-3956
Universidad de Granada (España)
Maribel Tercedor Sánchez
https://orcid.org/0000-0001-5390-5469
Universidad de Granada (España)

Recibido: 30 de septiembre de 2024
Aceptado: 18 de noviembre de 2025
Publicado: 27 de febrero de 2026

ABSTRACT

This article analyses the weaknesses and risks of neural machine translation (NMT) in clinical contexts where doctor-patient communication is relevant. It does so by focusing on the importance of the translation and adaptation of self-diagnostic questionnaires. The study evaluates the quality of the machine translation generated by Google Translate and DeepL applying an error analysis based on the Multidimensional Quality Metrics (MQM). The analysis of the NMT of the ChEAT questionnaire identifies several limitations in purely linguistic terms, and in pragmatic and sociocultural terms, which result in a lack of adaptation to the target audience, namely children. These limitations affect the quality of communication in medical settings, and stem from the NMT's inability to approach all the elements of the translation process from the outset, which ultimately leads to inaccurate results and perpetuated biases. The translation of scales is a multi-step process aimed at quality assurance. The need for human intervention in all phases of the translation process in clinical contexts and the danger of replacing human translation with unsupervised machine translation is therefore underlined.

KEYWORDS: machine translation; questionnaires; typology of errors; doctor-patient communication.

RESUMEN

Este artículo analiza los puntos débiles y los riesgos de la traducción automática neuronal (TAN) en contextos clínicos en los que la comunicación médico-paciente es relevante, centrándose en la importancia de la traducción y la adaptación de cuestionarios de autodiagnóstico. El estudio evalúa la calidad de la traducción generada por Google Translate y DeepL aplicando un análisis de errores basado en el marco Multidimensional Quality Metrics (MQM). Se han identificado errores lingüísticos, pragmáticos y socioculturales que comprometen la adecuación del texto a la población destinataria, los niños. Estas limitaciones afectan a la calidad de la comunicación en entornos médicos y derivan de la incapacidad de la traducción automática de abordar todas las fases del proceso de traducción, lo que lleva a resultados imprecisos y la perpetuación de sesgos. La traducción de cuestionarios es un proceso con múltiples etapas que garantizan el aseguramiento de la calidad. Por ello, este trabajo subraya la necesidad de la intervención humana desde el inicio del proceso y advierte de los riesgos de depender de la automatización sin supervisión humana en este ámbito.

PALABRAS CLAVE: traducción automática; cuestionarios; tipología de errores; comunicación médico-paciente.

1. Introducción

Los trastornos de la conducta alimentaria (TCA) representan un problema creciente de salud pública, con un aumento preocupante de su prevalencia entre niños y adolescentes (Noriega, 2023; Quelle y García, 2023; Sociedad Española de Médicos Generales y de Familia, 2018). Dada la importancia de la detección temprana de estos trastornos para evitar mayores complicaciones, tanto físicas como psicológicas, las escalas de diagnóstico se erigen como herramienta clave para evaluar síntomas, medir conductas de riesgo y facilitar el seguimiento y la comunicación entre médicos y pacientes. Estos cuestionarios, generalmente de autocumplimentación, no solo son relevantes en el ámbito clínico, sino que también contribuyen a la investigación y difusión de conocimiento científico a nivel internacional (Monterrosa-Castro et al., 2012).

Uno de los cuestionarios más utilizados para la detección de TCA en niños es el *Children's Eating Attitudes Test (ChEAT)*, diseñado específicamente para evaluar conductas alimentarias de riesgo en la infancia. No obstante, su eficacia depende en gran medida de su correcta compresión por parte del público infantil (Riley, 2004). Esto pone de manifiesto la importancia de la traducción humana en este ámbito, ya que, gracias a la traducción de cuestionarios, estas herramientas y sus resultados están disponibles en distintas lenguas, y gracias a su adaptación, sus contenidos resultan comprensibles para el público objetivo. Esta disponibilidad facilita el diagnóstico y tratamiento de los TCA en entornos culturales diferentes y favorece la diseminación de nuevos hallazgos al resto de la comunidad científica (Pérez Fernández y Miaja Menéndez, 2020). En este contexto, la traducción automática neuronal (TAN) ha experimentado un crecimiento sin precedentes. En los últimos años, la TAN se ha convertido en una herramienta eficiente, rápida y, sobre todo, ampliamente accesible para el público general (Andújar Carracedo, 2021; Savoldi et al., 2025). Esta accesibilidad ha facilitado su incorporación en múltiples ámbitos de la vida cotidiana, como el trabajo, los estudios, la comunicación personal y la comprensión básica de textos en otras lenguas. Un estudio realizado en Lituania por Kasperė et al. (2021) confirma esta tendencia, al documentar un uso generalizado de la TA entre distintos grupos sociales. Si bien la mayoría de los encuestados reconoce su utilidad, los resultados indican que aquellos con una mayor formación tienden a revisar y modificar los textos generados

por estas herramientas. El estudio concluye que resulta fundamental fomentar un uso crítico y consciente de la TA, especialmente en textos de alto riesgo. En este sentido, el uso no supervisado de traducciones automáticas en ámbitos especializados, como el médico, plantea riesgos significativos. Por ejemplo, aunque los cuestionarios de salud requieren procesos de traducción rigurosos y validados (Congost Maestre, 2010, 2016; Pérez Fernández y Miaja Menéndez, 2020; Sousa y Rojjanasrirat, 2011), el acceso libre a herramientas de TA permite que usuarios sin formación especializada recurran a versiones no revisadas para actividades como el autodiagnóstico. Esta práctica resulta especialmente preocupante, ya que errores de traducción en contextos clínicos pueden generar malentendidos con consecuencias graves para la salud. A pesar de los avances tecnológicos en el campo de la TA, diversos autores subrayan que la supervisión humana sigue siendo indispensable (Chen, 2024; Mohamed et al., 2021) particularmente en contextos de alto riesgo como el sanitario, donde las herramientas actuales aún presentan limitaciones para manejar con precisión los matices lingüísticos, pragmáticos y culturales presentes en los textos médicos (Baltabay, 2023; Díaz Prieto, 2012; He et al., 2020). Mehandru et al. (2022) añaden la necesidad de entrenar las herramientas con oraciones médicas previamente validadas, contemplar diferencias dialectales y niveles de alfabetización, y someter estas herramientas a evaluaciones rigurosas y advierten de que la implementacion de estas herramientas en contextos clínicos debe hacerse con responsabilidad, transparencia y siempre en función de la calidad del cuidado y la equidad en el acceso a la salud. Asimismo, la TAN puede perpetuar y crear sesgos de género y prejuicios socioculturales que pueden distorsionar el significado original del cuestionario y afectar en última instancia a la interpretación de sus respuestas (Savoldi et al., 2021).

1.1. Hipótesis y objetivos

El presente estudio se fundamenta en la hipótesis de que la TAN de acceso por parte del público general, aunque útil y accesible para traducciones rápidas, adolece de la precisión requerida para la traducción de textos clínicos sensibles como el ChEAT. Se postula que las imprecisiones lingüísticas, pragmáticas y culturales inherentes a la TAN, sumadas a su propensión a perpetuar sesgos de género, podrían comprometer la validez y adecuación de dicho cuestionario para su administración al público infantil. En consecuencia, este trabajo se plantea la siguiente pregunta de investigación: ¿en qué medida la traducción automática neuronal (TAN) de acceso por parte

del público general logra generar textos adecuados para niños, manteniendo la validez y fiabilidad del cuestionario original?

Para abordar esta pregunta, se propone un análisis exhaustivo de los errores más frecuentes observados en la TAN del cuestionario ChEAT, específicamente mediante el uso de Google Translate y DeepL, los motores de traducción más populares para el acceso por parte del público general. Se hará especial hincapié en el impacto de dichos errores sobre la adecuación y precisión del cuestionario, así como en la forma en que las marcas de género en las traducciones de la TAN puede perpetuar sesgos sexistas, en consonancia con lo observado por Savoldi et al. (2021). Este estudio no plantea abordar la calidad de herramientas de traducción automática neuronal adecuadas a géneros textuales y dominios de especialidad específicos, donde las labores de preedicion y posedición, el entrenamiento humano y la selección de corpus de calidad pueden garantizar resultados favorables, sino explorar cómo la naturaleza concisa y la sensibilidad del contenido de cuestionarios como el ChEAT hacen que los motores de acceso público de TAN resulten inadecuados. Se argumenta que la inversión de tiempo y recursos necesarios para una posedición completa superaría el esfuerzo requerido para una traducción realizada íntegramente por un traductor humano experto, dadas las particularidades lingüísticas y socioculturales del público infantil.

En consecuencia, se plantea como objetivo general evaluar la calidad de la traducción automática neuronal del cuestionario ChEAT, identificando y clasificando los errores más frecuentes. Para ello, se establecen los siguientes objetivos específicos:

1. Analizar las imprecisiones lingüísticas, pragmáticas y culturales presentes en la traducción automática del ChEAT.
2. Identificar y clasificar los sesgos de género perpetuados por la traducción automática del ChEAT.
3. Demostrar las limitaciones de la TAN de libre acceso para traducir adecuadamente textos clínicos sensibles dirigidos a público infantil, en los que la intervención humana es clave.
4. Evaluar en qué medida la traducción humana aporta matices a los cuestionarios que no están presentes en los resultados de la TAN de libre acceso.

2. Trastornos de la conducta alimentaria y traducción de herramientas de evaluación

2.1. TCA: prevalencia y grupos poblacionales afectados en España

Los trastornos de la conducta alimentaria (TCA) constituyen una variedad de trastornos mentales que se caracterizan por una alteración del comportamiento alimentario y de la relación de una persona con la comida y su cuerpo, afectando a la salud física y mental (Castañeda Quirama, 2020).

Las tasas de prevalencia de los TCA varían según el país, el grupo poblacional estudiado, el tipo de trastorno y su especificidad y las herramientas de evaluación (Arija Val et al., 2022). En España, es difícil establecer una tasa de prevalencia contrastada y demostrada debido a la heterogeneidad de los resultados y a la falta de consenso en cuanto a las herramientas diagnósticas utilizadas (Moreno Redondo et al., 2019). Por otro lado, es difícil acceder a datos precisos sobre el número de personas afectadas por un TCA que permitan estimar de manera realista el alcance de esta enfermedad, ya que son muchas las personas que no buscan tratamiento o que nunca han recibido un diagnóstico, ya sea por desconocimiento, negación o dificultad de acceso (Asociación TCA Aragon, 2020).

En 2018, la Sociedad Española de Médicos Generales y de Familia situó la prevalencia de los TCA en 4,1 % a 6,4 % en mujeres entre 12 y 21 años, y en 0,3 % para los hombres. De la misma forma, estableció que los TCA son la tercera enfermedad crónica más común entre adolescentes y alertó de la importancia de su prevención y diagnóstico precoz (Sociedad Española de Médicos Generales y de Familia, 2018). Otros estudios realizados posteriormente resaltan una tendencia acelerada al aumento desde la pandemia del COVID-19, lo que sugiere que el estrés, el aislamiento social y otros factores relacionados estrechamente con la crisis sanitaria exacerbaron los problemas de salud mental en la población general (Fernández-Aranda et al., 2020).

La mayoría de los estudios que abordan la traducción y adaptación de cuestionarios utilizan, predominantemente, muestras compuestas solo por mujeres (Montoro Pérez et al., 2024). Teniendo en cuenta que el grupo más afectado son las mujeres jóvenes, adolescentes y niñas, las herramientas de diagnóstico disponibles deben adaptarse para que se ajusten más a la

realidad cultural española y reflejen más fielmente las experiencias de este grupo poblacional. Sin embargo, no se puede pasar por alto que la edad no es un factor determinante para padecer la enfermedad y que los hombres también pueden sufrir TCA (Salas et al., 2011). Al revisar estudios recientes, Murray et al. (2017) subrayan que los TCA se han registrado por igual en hombres y mujeres desde el principio y señalan que existen diferencias en la manifestación de los síntomas entre ambos sexos. Por ejemplo, mientras que la naturaleza de la restricción alimentaria en la anorexia nerviosa femenina se orienta hacia la delgadez, los hombres tienden a buscar una apariencia delgada y musculosa (Schultz et al., 2022). Por lo tanto, aunque es crucial centrarse en los grupos más afectados, también es de vital importancia no excluir a otros grupos de población en el proceso de diagnóstico y tratamiento de estos trastornos.

Por otro lado, la bibliografía señala un aumento en la incidencia de TCA en niñas más jóvenes. La edad de inicio de conductas relacionadas con los TCA es cada vez menor. Son cada vez más los casos detectados en edades tempranas y diversos medios y asociaciones se están haciendo eco de esta tendencia (Noriega, 2023; Quelle y García, 2023). Este aumento puede estar relacionado con la influencia de los medios de comunicación y los inalcanzables ideales de belleza que promueven (Behar, 2010). En cualquier caso, este fenómeno es especialmente preocupante por las implicaciones que tiene para la salud física y emocional de los niños, que carecen de las herramientas de gestión emocional de las que, por experiencia vital, puede disponer la población adulta. Precisamente, distintos estudios señalan la falta de herramientas de evaluación específicas para niños que se adapten a su edad y capacidades cognitivas (Riley, 2004).

2.2. Escalas de autodiagnóstico de TCA: definición e instrumentos más utilizados

Las escalas son las herramientas de evaluación más utilizadas en el ámbito de la salud, pues ayudan a conocer la situación de los pacientes, detectar problemas de salud y facilitar su seguimiento. En el ámbito de la psicología y de los TCA, se utilizan escalas de autodiagnóstico, es decir, cuestionarios de autocumplimentación. Estas herramientas están diseñadas para evaluar conductas o síntomas que no se pueden detectar a través de pruebas físicas, pero que pueden estar relacionados con el desarrollo de un TCA (Monterrosa-Castro et al., 2012). Por lo tanto, el papel de estas escalas al comienzo de la evaluación psicológica es de una importancia

crítica (Bolaños Ríos, 2013). Así, tanto profesionales como afectados pueden obtener una primera impresión, en ocasiones condicionando la decisión posterior de realizar más pruebas que confirmen tal diagnóstico.

Las escalas están compuestas por ítems en forma de afirmaciones sobre la conducta alimentaria, que se califican en función del grado de identificación. El fin de estos elementos es definir la puntuación de cada una de las categorías y matizar la evaluación de conductas de riesgo (Morales et al., 2020). La *Guía de práctica clínica sobre trastornos de la conducta alimentaria*, publicada por el Ministerio de Sanidad y Consumo de España en 2009, recoge los cuestionarios más utilizados. No obstante, destaca por su uso el *Eating Attitudes Test* (Garner y Garfinkel, 1979), en su versión de 40 preguntas (EAT-40) , así como en su versión de 26 preguntas (EAT-26) (Garner et al., 1982) y en su versión adaptada al público infantil (ChEAT) (Maloney et al., 1988).

2.3. Traducción de escalas: proceso de retrotraducción

En su mayoría, las herramientas de autodiagnóstico se redactan originalmente en inglés, que actúa como lengua de difusión científica internacional (Congost Maestre, 2010). Son varias las razones por las que se opta por la traducción de escalas de autodiagnóstico, en lugar de la elaboración de nuevas. En primer lugar, la traducción permite a los investigadores y miembros de la comunidad científica acceder a instrumentos ya validados, con propiedades psicométricas ya reconocidas y con una base empírica sólida. Así, sus resultados pueden ser compartidos entre diferentes estudios, enriqueciendo de esta manera el conocimiento sobre este tipo de trastornos a nivel global. De igual manera, elaborar escalas nuevas y todo el proceso de validación que esto implica, resulta mucho más costoso que traducir escalas ya validadas (Pérez Fernández y Miaja Menéndez, 2020).

No obstante, la traducción y adaptación cultural de escalas de autodiagnóstico y su subsiguiente validación conforman un proceso complejo que requiere de la misma atención que la creación de una nueva escala. Sousa y Rojjanasrirat (2011) describen el proceso en varios pasos. En primer lugar, se realiza una traducción inicial llevada a cabo por al menos dos traductores experimentados. A continuación, se comparan ambas versiones y consensua una primera traducción preliminar. Esta versión es traducida de nuevo al idioma original en un

proceso de *blind back-translation* o retrotraducción ciega. Un comité multidisciplinar se encarga de comparar estas dos retroversiones con la escala original para acabar con una versión prefinal con la que realizar una prueba piloto que recoja una muestra de participantes con características similares al público al que está destinado la escala. Por último, se llevan cabo pruebas psicométricas que confirmen su validez y fiabilidad.

2.4. Directrices actuales en torno a la elaboración de cuestionarios

Dentro de esta cadena, hay que destacar el previo y necesario proceso de documentación que conlleva esta especialidad textual. En cuanto a las directrices en torno al formato de preguntas y respuestas, Congost Maestre (2012) advierte de la necesidad de establecer preguntas y respuestas bien delimitadas y que no se presten a distintas interpretaciones. Las alternativas deben excluirse mutuamente con el fin de despejar las posibles dudas y facilitar la elección de la respuesta más adecuada.

En el capítulo *Diseño de cuestionarios* del libro *Métodos de Investigación clínica y epidemiológica*, Argimón Pallás y Jiménez Villa (2004) recomiendan evitar preguntas ambiguas y huir de formulaciones vagas como «en ocasiones», o «a menudo».

En cuanto al uso de la lengua, la literatura indica un estilo llano y claro, lo cual pasa por hacer uso de sintaxis sencilla, frases cortas, voz activa y tiempos verbales simples. Asimismo, concluye que los aspectos lingüísticos que con mayor frecuencia suelen dar problema a la hora de traducir son los siguientes: polisemia léxica, expresiones coloquiales, orden de secuenciación de términos, repetición léxica, grado de formalidad, registro léxico, falsos amigos, concisión del inglés, gradación de las respuestas y reducción léxica (Congost Maestre, 2016: 132-133).

2.5. Traducción automática: evaluación de la calidad y errores más comunes

Con la llegada reciente de la traducción automática neuronal (TAN), el rendimiento de los sistemas de TA ha mejorado sustancialmente, pues se pueden entrenar para una sensibilidad al contexto hasta hace no tantos años impensable (Andújar Carracedo, 2021; Rivera-Trigueros, 2022). Sin embargo, aunque la TAN puede ser efectiva en determinados contextos (Koehn, 2020), su rendimiento varía en función de múltiples factores, como el par de lenguas, el

entrenamiento del modelo y el tipo de texto. En contextos especializados, como el ámbito clínico, la calidad de la traducción depende en gran medida del uso de motores entrenados específicamente en terminología médica y con textos de calidad. Sin embargo, precisamente en contextos específicamente clínicos, hay factores pragmáticos y culturales que escapan a la tecnología actual, matices que son particularmente importantes en cuestionarios, donde la forma en la que se formula una pregunta puede influir en la respuesta del paciente y provocar problemas de comunicación entre pacientes y médicos, perjudicando así la calidad de la atención que recibe el paciente en su diagnóstico (Mehandru et al., 2022).

La evaluación de los sistemas de TAN es un tema complejo que involucra la interacción de factores lingüísticos y extralingüísticos. Esta complejidad se traduce en una falta de consenso claro y de estandarización en los criterios de evaluación, lo que se debe a la naturaleza multifacética del proceso, algo que ha llevado a discrepancias significativas en las expectativas de calidad. En la industria, la calidad se define en función de la satisfacción del cliente, que no siempre coincide con el usuario final. Desde la perspectiva académica, la evaluación de calidad se enfoca en enfoques lingüísticos y métricas estructuradas (Castilho et al., 2018). Rivera-Trigueros (2021) considera por su parte que la calidad se mide a partir del esfuerzo humano de posedición y subraya la necesidad de comprender el alcance completo de la intervención humana en el proceso de traducción. Sin embargo, ambos enfoques —el industrial y el académico— pueden ser complementarios. Una traducción automática puede resultar ineficaz si requiere un excesivo tiempo y esfuerzo cognitivo por parte del traductor para lograr un producto final adecuado y adaptado a las necesidades del cliente (Rivera-Trigueros, 2021).

Castilho et al. (2018) señalan que la posedición ha de ser completa —en lugar de ligera— en textos de alta visibilidad o sensibilidad con calidad publicable, en este caso administrable, y que el esfuerzo de posedición puede evaluarse en función de tres criterios: esfuerzo temporal, esfuerzo técnico y esfuerzo cognitivo. Con ello se plantea medir si la posedición es una alternativa viable en comparación con la traducción humana desde cero, especialmente en función del tipo de texto que se esté tratando. Sin embargo, cabe destacar que el objetivo de esta investigación no es evaluar la viabilidad de la posedición, sino establecer una clasificación de errores en los que suele incurrir la TAN que permitirá reconocer patrones recurrentes que pueden afectar a la calidad de la traducción y facilitar una evaluación más precisa de posibles

necesidades de tareas de preedición, así como del esfuerzo requerido para la posedición.

Diversas investigaciones han tratado de identificar los puntos críticos en los que la traducción automática suele fallar. He et al. (2020) analizan la capacidad de la TAN para razonar con base a conocimiento extralingüístico, específicamente en la traducción del chino al inglés. El estudio se centra en evaluar cómo los modelos manejan ambigüedades léxicas y sintácticas que requieren sentido común para su correcta interpretación mediante un conjunto de pruebas con 1.200 ejemplos que evalúan tres tipos de ambigüedad: léxica, sintáctica sin contexto y sintáctica con contexto, que Díaz Prieto (2012) clasificó como léxica (palabras polisémicas) y estructural (varias interpretaciones de una misma oración). Los sistemas evaluados incluyen motores de acceso abierto y modelos ampliamente utilizados en investigación, como los basados en la arquitectura *Transformer* y *RNNSearch*, además de modelos de lenguaje preentrenados como *BERT* y *GPT-2*, que si bien no son motores de traducción automática *per se*, pueden contribuir al razonamiento contextual. Los hallazgos de este estudio subrayan que los modelos de TAN actuales tienen un desempeño deficiente en razonamiento con sentido común, con una precisión menor al 60,1 % y una consistencia inferior al 31 %, siendo las ambigüedades léxicas y sintácticas contextuales las más desafiantes. Además, se observó que aumentar el tamaño del corpus de entrenamiento mejora parcialmente la capacidad de razonamiento, aunque no soluciona completamente el problema. Finalmente, identificaron que más del 71 % de los errores en traducción están relacionados con fallos en el uso del sentido común, una habilidad inherentemente humana.

Al Sharou y Specia (2022) analizan los errores críticos en la TA en los pares de lenguas inglés-chino, inglés-italiano e inglés-árabe, con especial atención al contenido generado por usuarios (CGU). Su investigación se basa en la evaluación de traducciones producidas por motores comerciales y de acceso libre de amplio uso, tales como Google Translate, Bing y Systran. Su estudio propone y valida una taxonomía ampliada de errores críticos, que abarca dimensiones como la toxicidad del lenguaje, la presencia de riesgos para la salud y la seguridad, el tratamiento de nombres propios, la gestión de la negación, la exactitud en la traducción de numerales y fechas, la fidelidad en la transmisión de instrucciones, así como otras alteraciones del significado original. Sus hallazgos subrayan que los errores críticos no son infrecuentes ni exclusivos de un idioma y que ciertos elementos del texto fuente, tales como el lenguaje ofensivo, signos especiales, puntuación incorrecta y omisión de pronombres, aumentan la probabilidad de

errores graves en la traducción. Los autores concluyen que los sistemas de TA contemporáneos no abordan de manera adecuada estas problemáticas y que los esfuerzos deben dirigirse hacia la selección de los textos con los que los modelos se alimentan y el entrenamiento en el análisis de aspectos problemáticos.

Además, como se ha expuesto con anterioridad, la traducción automática puede exacerbar y perpetuar sesgos de género preexistentes. Siguiendo la taxonomía propuesta por Savoldi et al. (2021), estos sesgos se clasifican en dos categorías principales: sesgos representacionales, que implican la infrarrepresentación de ciertos grupos o el refuerzo de estereotipos, y sesgos de asignación, que resultan en una calidad de traducción desigual entre diferentes grupos, afectando la comprensión y relevancia del cuestionario para algunas personas. En el contexto específico de los cuestionarios sobre TCA, la asociación culturalmente prevalente de estos trastornos con el género femenino puede verse inadvertidamente amplificada. El estudio de Savoldi et al. (2021) identifica tres fuentes primordiales de sesgo: el sesgo preexistente, arraigado en actitudes y prácticas sociales; el sesgo técnico, derivado de limitaciones técnicas y decisiones inherentes al desarrollo del modelo; y el sesgo emergente, que surge de la interacción entre los sistemas y los usuarios. En el ámbito de los TCA, esto se podría traducir en traducciones que refuerzan estereotipos en torno a la imagen corporal femenina, que invisibilizan la prevalencia de estos trastornos en otros géneros o que emplean un lenguaje menos inclusivo, accesible o pertinente para hombres o personas no binarias.

Tras la revisión de algunos de los problemas recurrentes en la TA en general, y en la TAN en particular, se considera la noción de evaluación de la calidad en traducción, un objetivo central del presente trabajo, mediante la aplicación del marco de evaluación *Multidimensional Quality Metrics* (*MQM*). El *MQM* constituye un sistema ampliamente utilizado para la evaluación de la calidad en la traducción, ya sea humana, automática o generada por inteligencia artificial. Su modelo manual proporciona un marco detallado para identificar y clasificar errores en diversos niveles, lo que facilita una evaluación más precisa del rendimiento de un sistema de traducción. La tipología de errores en *MQM* se organiza en siete dimensiones principales: terminología, exactitud, convenciones lingüísticas (fluidez), estilo, convenciones locales, adecuación al público (veracidad) y diseño y lenguaje de marcado. No obstante, con el propósito de facilitar su aplicación en una variedad de contextos, se ha desarrollado una versión simplificada conocida

como *MQM-Core*, que reduce la complejidad de la taxonomía, conservando al mismo tiempo los elementos esenciales para la evaluación de la calidad. Las personas interesadas en emplear el *MQM* en sus propios análisis pueden acceder a la hoja de cálculo disponible en el sitio web del proyecto.

3. Metodología

Este estudio adopta un enfoque analítico y comparativo con el propósito de evaluar los errores de traducción de la TAN y el grado de precisión semántica de los cuestionarios de autodiagnóstico de TCA. Se eligió el *Children's Eating Attitudes Test (ChEAT)* como el principal objeto de estudio por estar dirigido a un público infantil y por ser este el cuestionario más utilizado. Para la comparación, se utilizó la versión original en inglés del *ChEAT*, así como las traducciones al español realizadas por DeepL y Google Translate.

Se optó por utilizar las versiones gratuitas de DeepL y Google Translate por varios motivos metodológicos y de accesibilidad. Ambas herramientas son ampliamente utilizadas por profesionales y público general, lo que las convierte en opciones relevantes para este análisis. Dado que el objetivo del estudio es evaluar el rendimiento de la traducción automática neuronal (TAN) en la traducción de cuestionarios clínicos sin intervención humana previa, resulta pertinente examinar los motores de TA a los que el público general tiene acceso inmediato. Los motores de TA de pago, como DeepL Pro, requieren licencias. Aunque existen sistemas de TA personalizados para contextos médicos, estos suelen estar restringidos a instituciones con acuerdos comerciales. La traducción de cuestionarios clínicos mediante herramientas de acceso libre es una práctica común entre el público general, especialmente en entornos como la información para consumo personal, donde la contratación de traductores profesionales no es viable. Este estudio busca analizar cómo funcionan estos motores en un entorno de acceso libre, y aportar una base de referencia para futuras investigaciones que puedan, por ejemplo, comparar el rendimiento de motores gratuitos con sistemas especializados en traducción médica. Los resultados obtenidos pueden servir como argumento para subrayar la importancia de la utilización de TA con intervención humana o para promover el desarrollo de motores más adecuados para textos clínicos. La evaluación de los errores y sesgos presentes en versiones no especializadas permite comprender los riesgos asociados a su uso sin posedición humana.

En primer lugar, se llevó a cabo un análisis del cuestionario en su versión original, con el fin de anticipar potenciales problemas que podrían surgir en el proceso de traducción automática. Además, se llevó a cabo una traducción humana inicial —a la que nos referimos como «traducción piloto»—, que sirvió como base para la comparación.

En segundo lugar, se llevaron a cabo las traducciones automáticas del ChEAT mediante DeepL y Google Translate, y los resultados se compararon con los obtenidos de del análisis inicial y la traducción humana. La comparación se presenta en una tabla que se incluye como Anexo 1. De esta manera, se analizaron y clasificaron sus principales errores, con el fin de evaluar su impacto en la comunicación médico-paciente y proponer soluciones adecuadas al género textual y al tema de los TCA.

Se identificaron errores en las traducciones generadas por ambos sistemas de TAN. Dicha evaluación se llevó a cabo utilizando la taxonomía propuesta por el modelo *Multidimensional Quality Metrics (MQM)*. La elección de *MQM* como marco de referencia se justifica por su carácter integral y adaptable. A diferencia de otras metodologías de evaluación que pueden centrarse únicamente en aspectos específicos (como la precisión terminológica o la fluidez lingüística), el *MQM* ofrece una visión holística que abarca tanto los errores lingüísticos como los aspectos relacionados con la adecuación cultural, la adaptación al público objetivo y el diseño.

La evaluación se llevó a cabo sobre un total de 26 ítems del cuestionario. Cada ítem fue traducido por ambas herramientas y evaluado mediante la tipología *MQM*-Core. Los errores se clasificaron en las siguientes categorías: *terminology*, *accuracy*, *linguistic conventions*, *style*, y *audience appropriateness*. Esta última es la única que solamente se consideró como categoría principal y única, al no corresponder ninguna de sus subcategorías –*culture-specific reference* y *offensive*– con el enfoque principal de la investigación en este aspecto, es decir, la edad. Se prescindió, asimismo, de las categorías *locale conventions* y *design and markup*, ya que estas no competen a esta investigación y engloban aspectos más allá de los meramente lingüísticos. Cada error fue clasificado además según su nivel de importancia, siguiendo una escala estándar *MQM*: *minor*, *major*, o *critical*, con una penalización ponderada (1, 5 y 25 puntos respectivamente). Se atribuyó la categoría de *minor* a aquellos errores que no afectan a la comprensión general ni distorsionan el mensaje; *major* para errores que afectan parcialmente el sentido del texto,

dificultan su comprensión o introducen ambigüedad; y *critical* para errores graves que alteran o contradicen completamente el significado original, induciendo a interpretaciones erróneas o incluso peligrosas.

4. Resultados

4.1. Análisis comparativo de errores de TAN: DeepL y Google Translate

Los resultados muestran que Google Translate obtuvo un puntaje de calidad superior al de DeepL. DeepL presentó un total de 31 errores identificados, con una distribución de 20 errores leves, 7 errores graves y 4 errores críticos. Por su parte, Google Translate también presentó un total de 31 errores, distribuidos en 22 errores leves, 5 graves y 4 críticos. Así, el sistema de Google presentó una penalización total (APT) de 147 puntos, mientras que DeepL acumuló 155 puntos. Esto se traduce en un puntaje general de calidad (OQS) de 18,33 para Google frente a 15,76 para DeepL, en una escala de 0 a 100. Aunque la cantidad total de errores fue semejante en ambos casos, la diferencia radica en el impacto cualitativo de los mismos: DeepL *a*cumuló un mayor número de errores graves, mientras que Google generó más errores menores, lo que se tradujo en una penalización global menor. El hecho de que DeepL presente más errores graves sugiere que, cuando este motor falla, lo hace dentro del ámbito de la semántica.

El análisis tipológico de los errores permite identificar las áreas de mayor conflicto para ambos motores. En el caso de DeepL, los errores más frecuentes se concentraron en cuatro categorías principales: estilo (29,03 %), precisión (25,81 %), terminología (25,81 %) y adecuación al público (16,13 %). Dentro de la categoría de estilo, los subtipos más comunes fueron los errores de registro lingüístico (12,90 %), estilo no idiomático (9,68 %) y estilo poco natural o forzado (6,45 %). En cuanto a la precisión, el 16,13 % de los errores correspondió a traducciones incorrectas del significado original (*mistranslation*), mientras que el 9,68 % se debió a casos de *overtranslation* o sobretraducción. Si bien esta subcategoría puede abarcar distintos tipos de errores, en este análisis concreto solo se identificaron casos relacionados con sesgos de género, al considerarse que la especificación del género implica la inclusión de información no presente en el texto original. En la categoría de terminología, se identificaron tanto errores por uso inconsistente de términos (16,13 %) como por elección de términos incorrectos (9,09 %). Finalmente, se registró un único error de tipo gramatical (3,23 %), lo que indica una relativa

solidez en cuanto al cumplimiento de convenciones lingüísticas básicas.

En el caso de Google Translate, la mayoría de los errores también se concentraron en las categorías de estilo (35,48 %), terminología (22,58 %), precisión (19,35 %) y adecuación al público (16,13 %). Dentro de los errores de estilo, el subtipo más recurrente fue el estilo no idiomático, que representó el 16,13 % del total de errores, seguido por errores de registro (12,9 %) y estilo forzado o poco fluido (6,45 %). En lo que respecta a los errores terminológicos, el 16,13 % correspondió al uso inconsistente de términos y el 6,45 % al uso de términos incorrectos. Por su parte, los errores de precisión se dividieron entre traducción incorrecta del significado (9,68 %), sobretraducción (6,45 %) y subtraducción (3,23 %), esta última ausente en los resultados de DeepL. En este caso, se identificaron dos errores de tipo gramatical (6,45 %).

La predominancia de los errores de estilo en Google Translate sugiere una tendencia a generar construcciones lingüísticas que, si bien pueden ser gramaticalmente correctas, resultan poco idiomáticas, artificiales o alejadas del registro esperable, lo que sugiere una ligera ventaja para DeepL en la generación de estructuras más naturales y adaptadas al contexto. Por otro lado, los errores de precisión semántica que incluyen traducciones incorrectas del significado original, sobretraducción y subtraducción, fueron más frecuentes en DeepL, especialmente en la subcategoría de *mistranslation* (15,15 % en DeepL frente al 9,68 % en Google Translate). Este dato refuerza la idea de que DeepL, si bien en general mantiene una mayor fidelidad formal al texto fuente, tiende a errar en aspectos semánticos más profundos. A continuación, se ofrece un análisis cualitativo con el fin de analizar dichos resultados.

4.2. Adecuación al público: problemas de adaptación al destinatario

Uno de los aspectos más sensibles en la evaluación de traducciones del cuestionario ChEAT es el relativo a la adecuación del contenido al público infantil, categoría recogida explícitamente en la tipología del modelo *MQM*. Este parámetro no se refiere únicamente al uso correcto de la lengua desde una perspectiva formal, sino a la idoneidad del lenguaje, en ese caso, en función del grupo etario destinatario, su nivel de desarrollo cognitivo, su contexto sociocultural y la sensibilidad inherente a los temas tratados en el cuestionario, como la relación con la comida, el cuerpo y la autopercepción.

Antes de iniciar el proceso de traducción, se llevó a cabo un análisis exhaustivo del

instrumento original en inglés con el fin de identificar posibles elementos problemáticos, tanto desde el punto de vista terminológico como estructural. Este paso previo permitió anticipar dificultades asociadas a la complejidad semántica de ciertos ítems y planificar una documentación adecuada que facilitara una eventual adaptación. A pesar de ser una versión simplificada del EAT-26, el ChEAT conserva elementos que pueden dificultar su comprensión por parte de niños, como el uso de términos semiespecializados (*energy content*, *diet foods*), expresiones de connotación adulta (*to be dieting*) o construcciones que requieren un nivel de introspección y autoconciencia que no siempre está presente en menores (*I feel very guilty after eating*, expresión que presupone un desarrollo avanzado de la conciencia de culpa asociada a la conducta alimentaria).

Este contexto evidencia una limitación estructural de la TAN, tanto en DeepL como en Google Translate, que operan sin una adaptación real al tipo de receptor al que va dirigido el texto. Como resultado, las traducciones generadas por estos motores reproducen el contenido original en términos puramente lingüísticos, sin atender a las implicaciones cognitivas, emocionales o sociales de las palabras seleccionadas. Esta ausencia de matices contextuales se traduce en errores sistemáticos de adecuación al público, que afectan directamente la funcionalidad y la ética del cuestionario. Los resultados cuantitativos del análisis evidencian esta problemática. En ambas traducciones se identificaron cinco errores clasificados en la categoría de *audience appropriateness*, lo que representa un 16,13 % del total en DeepL y en Google Translate. Como era de esperar, estos errores se relacionan con el uso de tecnicismos o formulaciones abstractas inadecuadas para niños. Asimismo, algunas estructuras observadas en las traducciones requieren procesos de introspección o de autorregulación que exceden las capacidades de una parte del público infantil. En estos casos, el cuestionario deja de ser un instrumento de medición válido para convertirse en un texto que impone categorías conceptuales adultas sobre sujetos que aún no están en condiciones de comprenderlas o asumirlas. Este riesgo se ve acentuado cuando se emplea un lenguaje crudo o excesivamente explícito, que podría reforzar estereotipos corporales o inducir preocupaciones en menores que, de otro modo, no las habrían interiorizado.

Si bien no forma parte del objetivo principal de la presente investigación, se incluyó una comprobación adicional derivada del interés exploratorio por parte de las investigadoras. Esta

consistió en solicitar a ChatGPT, modelo generativo de lenguaje, la traducción y adaptación al español del cuestionario, previa provisión de una descripción general del instrumento y de su público destinatario. Dicha prueba se plantea como una observación complementaria, sin ánimo de desviar el foco del estudio, pero con el propósito de ofrecer una perspectiva adicional sobre el potencial actual de estas herramientas en tareas de traducción y adaptación textual. La traducción de ChatGPT se caracterizó por una adaptación mayoritariamente sintáctica. Algunas oraciones se simplificaron y acortaron, por ejemplo: *I give too much time and though to food* > «Pienso demasiado en la comida». También produjo secuencias explicativas para clarificar términos complejos, como «alimentos dietéticos», aunque la traducción resultante persistió en un nivel de complejidad inapropiado para el público infantil: *I eat diet foods* > «Como alimentos bajos en calorías». En general, las problemáticas inherentes a la utilización de términos semiespecializados y afirmaciones que instigan a la reflexión personal no fueron resueltas, sino meramente simplificadas: *I am aware of the energy (calorie) content in foods that I eat* > «Sé cuántas calorías tienen los alimentos que como». La simplificación sintáctica puede desempeñar un papel útil en la reducción de la carga cognitiva, favoreciendo así una lectura más fluida y una mayor concentración por parte de los niños, especialmente en edades tempranas. Como señalan Rebello et al. (2019) en tareas de lectura individual, la simplificación de estructuras gramaticales puede mejorar la comprensión de ideas centrales y detalladas, particularmente en lectores con experiencia limitada. No obstante, tal y como advierten Burchell et al. (2023) la comprensión lectora no depende únicamente de la estructura gramatical, sino también de la familiaridad con el vocabulario y del grado de desarrollo de habilidades lingüísticas y conceptuales previas. En el caso del ChEAT, aunque la simplificación sintáctica puede facilitar el procesamiento superficial del enunciado, no garantiza por sí misma que el contenido sea plenamente comprendido, especialmente si persisten términos o nociones abstractas, técnicas o semiespecializadas. Esto puede limitar la accesibilidad semántica del ítem, incluso si la forma gramatical resulta sencilla.

4.3. Clasificación de errores según el modelo MQM

4.3.1. Errores terminológicos: selección inadecuada y uso inconsistente de términos

Uno de los errores más frecuentes detectados en ambas herramientas se encuentra en el manejo de la terminología. En DeepL, los errores terminológicos representaron el 25,81 % del

total, y en Google Translate, el 22,58 %. Estos errores pueden dividirse en dos grandes grupos: elección de términos incorrectos y uso inconsistente de la terminología.

El primero de ellos se refiere a la elección de una palabra que, aunque lingüísticamente válida, no refleja con precisión el sentido técnico o conceptual requerido en el contexto original, afectando así a la fidelidad de la traducción. Un ejemplo claro de este fenómeno se encuentra en la traducción de *I enjoy trying new rich foods*, donde tanto DeepL como Google Translate optaron por «Me gusta probar nuevos alimentos ricos». Se trata de un error derivado de la polisemia léxica, es decir, de la incapacidad de los motores para seleccionar el significado contextualmente adecuado cuando una palabra admite múltiples interpretaciones. En esta formulación, el adjetivo *rich* fue traducido como «ricos», una equivalencia léxica genérica que omite el matiz nutricional presente en el original, donde *rich* hace referencia al alto contenido calórico o graso del alimento. Ambas traducciones desplazan el eje semántico hacia una interpretación gastronómica o gustativa, alejándose del contenido que realmente se pretende evaluar. De forma similar, en la frase *I have gone on eating binges where I feel that I might not be able to stop*, traducida por DeepL como «He tenido atracones en los que creo que no podré parar», se observa un error terminológico al emplear «creo» en lugar de «siento». El uso del verbo «creer» en este contexto introduce una valoración racional que no se corresponde con la naturaleza emocional del original, en el que *feel* expresa una vivencia subjetiva de falta de control. Este desplazamiento semántico implica no solo una pérdida de matiz, sino una modificación conceptual del enunciado evaluado.

Una peculiaridad adicional de este tipo de textos es su formato estructural como cuestionario de respuestas cerradas, lo que exige una especial atención a la traducción de las escalas de respuesta. En este contexto, la selección terminológica no solo debe ser semánticamente precisa, sino que debe respetar la jerarquía interna de las opciones de respuesta para preservar la coherencia y la funcionalidad del instrumento. Por ejemplo, DeepL traduce *Often*, *Sometimes* y *Rarely* como «Algunas veces», «A menudo» y «Usualmente». El orden de gradación no queda claro. En inglés, *Often* denota una alta frecuencia, *Sometimes* una frecuencia intermedia y *Rarely* una baja frecuencia. Sin embargo, en español, la elección de «Usualmente» como equivalente de *Rarely* resulta problemática y errónea, ya que este término sugiere una frecuencia alta en lugar de baja. Esta imprecisión en la escala de gradación puede generar ambigüedad en la

interpretación de los cuestionarios.

Otro aspecto relevante dentro de esta categoría es la inconsistencia terminológica detectada en la traducción del término *food/foods*, que alterna de manera arbitraria entre «comida» y «alimentos» en distintos ítems del cuestionario. En español, tanto «comida» como «alimentos» pueden ser traducciones válidas de *food*, dependiendo del contexto. Sin embargo, su uso no es completamente intercambiable en todos los contextos ni en todos los registros. El término «comida» es más frecuente en la comunicación cotidiana, tiene una carga más coloquial y cercana, y es ampliamente entendido. Por el contrario, «alimentos» se asocia a menudo con contextos más técnicos, normativos o vinculados a la nutrición formal. En este sentido, la elección alternante entre «comida» y «alimentos» por parte de los motores de traducción —como puede observarse en ítems traducidos por DeepL y Google Translate, donde *food* se convierte en «comida» en unos casos y *foods* en «alimentos» en otros— introduce una variación terminológica innecesaria y no motivada semánticamente. Desde un punto de vista traductológico, tratándose de un género en el que la variación puede confundir, resulta más coherente y estilísticamente uniforme optar por una sola forma.

4.3.2. Errores de precisión y gramaticales: distorsiones del sentido original

Una parte significativa de los errores detectados en las traducciones automáticas analizadas corresponde a la categoría de *Accuracy* del modelo MQM, en particular a las subcategorías *Mistranslation* y *Overtranslation*, siendo estos últimos considerados en su totalidad como casos de sesgo de género al tratarse de adiciones de información no especificada en el original, tal como se ha señalado previamente. En el caso de DeepL, los errores de traducción representaron el 16,13 % del total y los de sobretraducción, un 9,68 %. En Google Translate, la proporción fue ligeramente inferior: 9,68 % de errores de traducción y 6,45 % por sobretraducción, además de un 3,23 % adicional de errores de subtraducción, categoría no presente en los resultados de DeepL.

La subcategoría *Mistranslation* agrupa los errores en los que el contenido del texto traducido no refleja correctamente el significado del original, generando cambios semánticos relevantes. Un ejemplo especialmente ilustrativo aparece en la traducción del ítem *I have gone on eating binges where I feel that I might not be able to stop*. DeepL ofrece la versión «He tenido atracones

en los que creo que no podré parar». Aquí se sustituye *might not be able to stop* —que expresa posibilidad e incertidumbre— por «no podré parar», lo cual constituye una afirmación rotunda. Este desplazamiento elimina el carácter tentativo y emocional del original, transformándolo en una declaración racional y definitiva, lo que implica un cambio de significado sustancial. Otro ejemplo es el caso de la traducción por parte de Google Translate de *I can show self-control around food* como «Me controlo con respecto a la comida». Esta expresión tiene una connotación positiva de enorgullecimiento. No es lo mismo controlarse con respecto a la comida (una acción más limitada y puntual), que mostrar autocontrol (una acción más consciente y duradera). Otro caso particularmente significativo se encuentra en el ítem *I am scared about being overweight*, traducido por DeepL como «Me asusta el sobrepeso». Si bien esta formulación es gramaticalmente correcta, representa un cambio sustancial de significado respecto al original. Mientras que la versión inglesa expresa un temor personal y subjetivo vinculado a la posibilidad de tener sobrepeso, la traducción en español presenta el enunciado como un miedo generalizado hacia el concepto de «sobrepeso». Esta construcción cosifica y externaliza el objeto del miedo, y puede incluso interpretarse como una afirmación con connotaciones gordófobas, al convertir el sobrepeso en una entidad abstracta y amenazante, más allá de la experiencia individual del hablante. En lugar de reflejar una emoción interna relacionada con la autopercepción corporal, la traducción sugiere una postura de rechazo hacia una categoría social o física. Este tipo de errores suelen estar relacionados con el concepto de «ambigüedad estructural», la cual se debe a la falta de información sintáctica en una oración. Otro ejemplo del mismo problema es la traducción de *I think a lot about having fat on my body* como «Pienso mucho en tener grasa en el cuerpo», ya que, tal y como está formulada, no queda claro si la persona desea tener grasa en el cuerpo o no.

Por otro lado, los errores de *overtranslation* se producen cuando la traducción introduce información no presente en el original y, en este caso, los constituyen íntegramente errores por sesgo de género. Por ejemplo, DeepL traduce *I think a lot about wanting to be thinner* como «Pienso mucho en que quiero estar más delgado». Aquí se introduce una marca de género que no está presente en el texto fuente. De igual forma, en el ítem *Other people think I am too thin*, DeepL mantiene el masculino con «Los demás piensan que estoy demasiado delgado», mientras que Google Translate, por el contrario, opta por el femenino: «Otras personas piensan que estoy

demasiado delgada». Este fenómeno, aunque superficialmente pueda parecer un intento de adecuación cultural, constituye un caso de *overtranslation*, al tratarse de una interpretación adicional que no está explícitamente codificada en el original, y cuya inclusión tiene efectos discursivos significativos.

Dentro de este punto, no obstante, resulta relevante traer a primer plano la cuestión del lenguaje inclusivo y los sesgos de género implícitos en estas decisiones automáticas. DeepL tiende a emplear sistemáticamente el masculino genérico en todo el cuestionario, como se observa en ejemplos como «Pienso mucho en que quiero estar más delgado» o «Los demás piensan que estoy demasiado delgado». Por su parte, Google Translate presenta un comportamiento más irregular: en el caso anterior, introduce el femenino solo cuando se hace referencia directa a la delgadez, mientras que previamente ha traducido *I feel uncomfortable after eating sweets* como «Me siento incómodo después de comer dulces». Esta elección estadística, probablemente derivada del corpus de entrenamiento utilizado —donde la asociación entre delgadez y mujeres es más frecuente en el discurso sobre TCA—, puede tener consecuencias no deseadas. Desde el punto de vista cultural, podría interpretarse como una representación más realista de la prevalencia de los TCA en mujeres y niñas; sin embargo, no constituye una estrategia inclusiva, ya que excluye implícitamente a los varones de una realidad clínica que también les afecta, perpetuando estereotipos de género y alimentando el tabú y el estigma que aún rodea los TCA en población masculina. De la misma forma, la bibliografía sugiere que los síntomas de TCA no son los mismos para hombres y mujeres, por lo que debería buscarse la neutralidad. Børset (2021) menciona en su trabajo estrategias que velan por un uso inclusivo del lenguaje. Entre ellas, destaca la estrategia de neutralización a través del uso de palabras genéricas tales como adjetivos, pronombres y determinantes sin marca de género, sustantivos abstractos y sustantivos colectivos. Señala técnicas como la sustantivación de procesos y funciones de una persona y la opción de utilizar el sustantivo «persona» en vez del sustantivo «hombre». Dentro de la categoría de neutralización, y sobre la posibilidad de utilizar la «e» como terminación del género neutro, menciona diversas instituciones como el CEP-CIE, la UPV o la ONU, que están a favor de la introducción del género neutro con dicha terminación, pero recomiendan restringir su uso a ámbitos informales ya que podría presentar problemas gramaticales.

Por último, aunque menos frecuentes en términos cuantitativos, e incluidos en esta categoría por su impacto en la transmisión precisa del mensaje, se encuentran los errores gramaticales identificados dentro de la categoría de *Linguistic conventions*. El único caso identificado es el del ítem *I have gone on eating binges where I feel that I might not be able to stop*, cuya complejidad sintáctica y semántica fue gestionada de forma deficiente por ambos motores. En el original, la estructura combina un presente perfecto (*have gone on*) con un modo condicional (*might not be able to stop*), lo cual sitúa la acción en el pasado reciente y la matiza con un grado de incertidumbre emocional. En la versión ofrecida por DeepL —«He tenido atracones en los que creo que no podré parar»— se detecta un claro error gramatical al emplear una construcción en presente («creo que no podré») que desajusta la lógica temporal del original: convierte la experiencia introspectiva y emocional en una declaración racional y proyectiva hacia el futuro, además de no concordar con el verbo en pasado imperfecto. De forma similar, Google Translate ofrece «He tenido atracones de comida en los que siento que no podría parar», que, aunque mejora levemente la opción de verbo («siento» en lugar de «creo»), y el sentido de incertidumbre con el condicional, mantiene la falta de adecuación en el uso de tiempos verbales.

4.3.3. Errores de estilo: construcciones poco idiomáticas y registro inadecuado

Los errores de estilo constituyen una de las categorías más representativas del análisis tipológico, con una incidencia notable en ambos sistemas de traducción automática. En el caso de DeepL, los errores de esta categoría alcanzan el 29,03 % del total, mientras que en Google Translate ascienden al 35,48 %, lo que los posiciona como el grupo más numeroso. Dentro de esta macrocategoría, se identificaron errores recurrentes en tres subtipos principales del modelo MQM: *register*, *awkward style* y *unidiomatic style*.

Los errores de registro se producen cuando la elección de vocabulario o estructuras sintácticas no es coherente con el nivel de formalidad, familiaridad o contexto comunicativo que requiere el texto. Este tipo de desajuste fue especialmente visible en expresiones marcadas por un léxico técnico, ya comentado, o expresiones excesivamente formales que resultan inapropiadas a la naturalidad con la que debería formularse un cuestionario orientado a la autoevaluación de niños. Un caso ilustrativo aparece en la traducción del ítem *I am scared about being overweight*. DeepL ofrece la formulación «Me asusta el sobrepeso», mientras que Google

Translate propone «Me da miedo tener sobrepeso». Aunque ambas traducciones son gramaticalmente correctas, difieren en su grado de formalidad y naturalidad. El verbo «asustar» seguido de un sustantivo abstracto como *el sobrepeso* da lugar a una expresión de corte más técnico o institucional, que puede resultar distante para muchos hablantes. En cambio, «dar miedo» es una construcción más habitual en registros coloquiales y cercanos al público infantil. Otro ejemplo significativo lo constituye la traducción de *I give too much time and thought to food*. DeepL propone «Dedico demasiado tiempo y reflexión a la comida», mientras que Google Translate ofrece «Dedico demasiado tiempo y pensamiento a la comida». En ambas versiones, la elección léxica de «reflexión» o «pensamiento» responde a una traducción literal del término *thought*, pero introduce un nivel de abstracción y formalidad inusual en el ámbito infantil.

Los errores considerados como *awkward style* recogen estilos que presentan una verbosidad excesiva o estructuras sintácticas innecesariamente complejas, a menudo como resultado de una retención inadecuada del estilo del texto original en la lengua meta. Es decir, el problema no radica en la corrección gramatical del enunciado, sino en su artificiosidad o densidad poco adecuada para el contexto comunicativo. Un ejemplo claro se encuentra en el ítem I *have gone on eating binges where I feel that I might not be able to stop*, cuya versión de Google Translate es «He tenido atracones de comida en los que siento que no podría parar». La adición de «comida» es redundante, ya que el término atracón en español ya presupone la referencia al acto de comer. Esta sobreespecificación, además de innecesaria, produce una carga expresiva desproporcionada que entorpece la fluidez del texto. Otro ejemplo representativo aparece en *I give too much time and thought to food*, traducido como «Dedico demasiado tiempo y reflexión a la comida» o «Dedico demasiado tiempo y pensamiento a la comida», que reproducen literalmente la estructura del original sin adaptarla al patrón natural del español.

El error clasificado como *unidiomatic style* designa aquellas construcciones que, pese a ser gramaticalmente correctas, resultan antinaturales en la lengua de llegada. Es decir, son frases que un hablante nativo no produciría de forma espontánea, ya sea por combinaciones léxicas inusuales, estructuras sintácticas poco frecuentes o interferencias de la lengua original. Por ejemplo, en la traducción del ítem *I think a lot about wanting to be thinner*, DeepL ofrece la versión «Pienso mucho en que quiero estar más delgado», mientras que Google Translate propone «Pienso mucho en querer adelgazar». Ambas construcciones, aunque correctas desde

el punto de vista sintáctico, resultan poco idiomáticas en español. La primera incurre en una construcción pesada y redundante, mientras que la segunda, con la secuencia «pensar en querer», presenta una concatenación de verbos que no refleja el uso fluido del español. Un segundo ejemplo relevante es el ítem *I think a lot about having fat on my body*, traducido por ambos motores como «Pienso mucho en tener grasa en el cuerpo». Esta construcción presenta una combinación léxica poco natural en español, donde no es habitual hablar de «tener grasa» de manera neutral. La expresión resulta ambigua y puede interpretarse como una afirmación neutra, incluso positiva, en lugar de una preocupación o conflicto corporal, que es el sentido que transmite el original.

5. Conclusiones

El presente estudio ha proporcionado una taxonomía de los errores más frecuentes detectados en la TAN de libre acceso, específicamente en Google Translate y DeepL, aplicada a cuestionarios de autodiagnóstico. Si bien el análisis se ha centrado en este género textual y ámbito específico, los hallazgos sugieren la aplicabilidad de dicha clasificación a otros contextos especializados clínicos. Los resultados obtenidos ponen de manifiesto las limitaciones de estos motores de TAN, tanto en el plano lingüístico como en el sociocultural, evidenciando su incapacidad para abordar adecuadamente las características propias del público destinatario y su propensión a perpetuar determinados sesgos preexistentes.

Desde un punto de vista cuantitativo, ambos motores obtuvieron puntuaciones de calidad bajas según el modelo *MQM*. Esta constatación refuerza la importancia de la traducción humana y la necesidad de replantear el papel de la TA en textos dirigidos a pacientes, donde un resultado erróneo en la traducción podría comprometer su salud y bienestar, especialmente en contextos automatizados (O'Brien, 2023). La categorización de errores mostró que las áreas más problemáticas en ambas herramientas son el estilo, la precisión semántica y la terminología. Dichas limitaciones se derivan de la omisión, por parte de las herramientas, de fases fundamentales del proceso de traducción, reduciéndose a una mera transferencia probabilística del texto de una lengua a otra. Los datos analizados en este estudio confirman que, en textos de alta sensibilidad y finalidad diagnóstica, la traducción automática puede servir como herramienta de apoyo, pero no como solución autónoma ni fiable en términos de calidad

funcional o ética profesional. Para profundizar en esta línea de investigación, sería pertinente llevar a cabo estudios de recepción que evalúen el impacto de la traducción automática neuronal en las respuestas proporcionadas por los participantes. El objetivo de dichos estudios sería determinar si las categorías de problemas identificadas en este trabajo inducen, efectivamente, a respuestas erróneas.

Por otro lado, en este contexto particular de cuestionarios breves, pero de alta trascendencia, incluso la posedición completa de la traducción automática puede dejar matices fuera frente a una traducción humana completa. La naturaleza concisa y sensible de estos instrumentos y la trascendencia de su correcta interpretación exigen una atención meticulosa desde las primeras etapas del proceso de traducción, ya que pequeños cambios semánticos, elecciones léxicas inadecuadas o estructuras estilísticamente ajenas pueden alterar por completo la validez de un ítem. Esta complejidad se agrava cuando los segmentos se traducen de forma aislada y sin un contexto amplio, lo cual podría favorecer una dependencia acrítica de las soluciones generadas por la traducción automática. En este sentido, cabría plantear —como posible línea de investigación futura— si la exposición continuada a soluciones generadas por traducción automática en el contexto específico de textos fragmentarios, como cuestionarios compuestos por ítems breves e independientes, favorece una forma de acomodación pasiva del traductor humano —y, por ende, del usuario final— a formulaciones artificiales o poco idiomáticas. Dado que estos segmentos suelen carecer de un contexto discursivo más amplio que permita guiar la interpretación, la posedición podría realizarse de forma más acrítica, aceptando estructuras que, en otros entornos textuales más cohesionados, serían fácilmente detectadas como impropias o no naturales. Esta posible intoxicación estilística inducida por la TA en entornos fragmentarios plantea interrogantes relevantes sobre los efectos de la automatización en la competencia traductora, y constituye una línea de investigación aún poco explorada. La inversión de tiempo y recursos en una posedición exhaustiva podría incluso superar los esfuerzos requeridos en una traducción realizada íntegramente por un traductor humano experto.

En conclusión, la TAN al alcance del público general, ejemplificada en las versiones actuales de Google Translate y DeepL, requiere un profundo proceso de revisión por parte de traductores humanos, acompañado de una labor previa de documentación y análisis de las características

del texto y de sus implicaciones. Esta necesidad de supervisión humana reafirma la importancia de la experiencia y el criterio del traductor profesional para garantizar la calidad, precisión y pertinencia de las traducciones, especialmente en contextos sensibles como el de la salud, donde la minimización de errores es de vital importancia.

Bibliografía

Al Sharou, Khetam, y Specia, Lucia (2022). A Taxonomy and Study of Critical Errors in Machine Translation. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 171-180.

Andújar Carracedo, Ángel (2021). Traducción automática neuronal sensible al contexto [Trabajo fin de máster]. Universidade Politécnica de Valencia. https://riunet.upv.es:443/handle/10251/172540

Argimón Pallás, Josep María y Jiménez Villa, Josep (2004). Diseño de cuestionarios. En Josep Maria Argimon Pallàs y Josep Jiménez Villa, *Métodos de investigación clínica y epidemiológica* (3.ª ed., pp. 188-200). Elsevier.

Arija Val, V., Santi Cano, M. J., Novalbos Ruiz, J. P., Canals, J. y Rodríguez-Martín, A. (2022). Caracterización, epidemiología y tendencias de los trastornos de la conducta alimentaria. *Nutrición Hospitalaria,* 8-15. https://doi.org/http://dx.doi.org/10.20960/nh.04173

Asociación TCA Aragón. (2020). Estadísticas sobre los TCA. https://www.tca-aragon.org/2020/06/01/estadisticas-sobre-los-tca/

Baltabay, Dana (2023). *Peculiarities and Challenges of Machine Translation (MT): The Role of Machine Translation in the life of Translators* [M. S. Narikbaev KAZGUU University]. http://repository.kazguu.kz/handle/123456789/1635

Behar A, Rosa (2010). La construcción cultural del cuerpo: El paradigma de los trastornos de la conducta alimentaria. *Revista chilena de neuro-psiquiatría, 48*(4), 319-334. https://doi.org/10.4067/s0717-92272010000500007

Bolaños Ríos, Patricia (2013). Cuestionarios, Inventarios y Escalas. *Trastornos de la Conducta Alimentaria,* 18, 1981-2007. http://www.tcasevilla.com/archivos/cuestionarios,_inventarios_y_escalas_en_tca.pdf

Børset, Ingrid Kristine (2021). *¿Lenguaje que incluye o lenguaje que excluye? Ventajas y dificultades de las estrategias sobre lenguaje inclusivo de género en siete guías*. Universidad de Oslo.

Burchell, Diana, Hipfner-Boucher, Kathleen, Deacon, S. Hélène, Koh, Poh Wee y Chen, Xi (2023). Syntactic Awareness and Reading Comprehension in Emergent Bilingual Children. *Languages, 8*(1), 62. https://doi.org/10.3390/languages8010062

Castañeda Quirama, Tatiana (2020). Perfil clínico de pacientes con trastornos de la conducta alimentaria. *Journal of Science, Humanities and Arts, 7*(2). http://dx.doi.org/10.17160/josha.7.2.648

Castilho, S., Doherty, S., Gaspari, F.y Moorkens, J. (ed.) (2018). Approaches to Human and Machine Translation Quality Assessment. *Translation Quality Assessment: from principles to practice* (pp. 9-38). Springer.

https://doi.org/10.1007/978-3-319-91241-7_2

Chen, Yuduo (2024). The metamorphosis of machine translation: The rise of neural machine translation and its challenges. *Applied and Computational Engineering, 43*(1), 99-106. https://doi.org/10.54254/2755-2721/43/20230815

Congost Maestre, Nereida (2010). El lenguaje de las Ciencias de la Salud: Los cuestionarios de salud y calidad de vida y su traducción del inglés al español [Tesis doctoral]. Universidad de Alicante. http://rua.ua.es/dspace/bitstream/10045/17562/1/Tesis_congost.pdf

Congost Maestre, Nereida (2012). Aspectos formales y visuales en los cuestionarios de salud y calidad de vida. *Panace@, 13*(35), 99-112. https://rua.ua.es/dspace/bitstream/10045/34326/1/2012_Congost_Panacea.pdf

Congost Maestre, Nereida (2016). Aspectos lingüísticos en la traducción de cuestionarios de salud (británicos y estadounidenses). *The Journal of Specialised Translation,* 26, 116-135. http://hdl.handle.net/10045/57482

Díaz Prieto, Petra (2012). Luces y sombras en los 75 años de traducción automática. En Juan José Lanero y José Luis Chamosa (ed.), *Lengua, traducción, recepción: en honor de Julio César Santoyo* (Vol. 2, pp. 139-175). Universidad de León. http://hdl.handle.net/10612/4712

Fernández-Aranda, F., Casas, M., Claes, L., Clark Bryan, D., Favaro, A., Granero, R., Gudiol, C., Jiménez-Murcia, S., Karwautz, A., Le Grange, D., Menchón, J. M., Tchanturia, K. y Treasure, J. (2020). COVID-19 and implications for eating disorders. *European Eating Disorders Review, 28*(3), 239-245. https://doi.org/10.1002/erv.2738

Garner, David M., y Garfinkel, Paul E. (1979). The Eating Attitudes Test: an index of the symptoms of anorexia nervosa. *Psychological medicine, 9*(2), 273-279. https://doi.org/10.1017/S0033291700030762

Garner, D. M., Olmsted, M. P., Bohr, Y. y Garfinkel, P. E. (1982). The Eating Attitudes Test: Psychometric features and clinical correlates. *Psychological Medicine*, 12, 871-878. https://doi.org/10.1017/s0033291700049163v

He, J., Wang, T., Xiong, D. y Liu, Q. (2020). The Box is in the Pen: Evaluating Commonsense Reasoning in Neural Machine Translation. En *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3662-3672). Association for Computational Linguistics.

Koehn, Philip. (ed.). (2020). Current Challenges. En *Neural Machine Translation* (pp. 293-310). Cambridge University Press. https://doi.org/10.1017/9781108608480.017

Maloney, Michael J., McGuire, Julie Bell y Daniels, Stephen R. (1988). Reliability testing of a children's version of the Eating Attitude Test. *Journal of the American Academy of Child and Adolescent Psychiatry,* 27, 541-543. https://doi.org/10.1097/00004583-198809000-00004

Mehandru, Nikita, Robertson, Samantha y Salehi, Niloufar (2022). Reliable and Safe Use of Machine Translation in Medical Settings. En *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2016-2025). ACM. https://doi.org/10.1145/3531146.3533244

Ministerio de Sanidad y Consumo. (2009). *Guía de Práctica Clínica sobre Trastornos de la Conducta Alimentaria*. https://www.aeesme.org/wp-content/uploads/docs/GPC_440_TCA_Cataluya.pdfv

Mohamed, Shereen A., Elsayed, Ashraf A., Hassan, Y. F., y Abdou, Mohamed A. (2021). Neural machine

translation: past, present, and future. *Neural Computing and Applications*, 33, 15919-15931. https://doi.org/10.1007/s00521-021-06268-0

Monterrosa-Castro, Álvaro, Boneu-Yépez Deiby, John, Muñoz-Méndez, José Tomás y Almanza-Obredor, Pedro Enrique. (2012). Trastornos del comportamiento alimentario: escalas para valorar sintomas y conductas de riesgo. *Revista Ciencias Biomédicas,* 3, 99-111. https://doi.org/10.32997/rcb-2012-3173

Montoro Pérez, Néstor, Montejano Lozoya, Raimunda, Martín Baena, David, Talavera Ortega, Marta y Gómez Romero, María Rosario (2024). Propiedades psicométricas del Eating Attitudes Test-26 en escolares españoles. *Anales de Pediatría*, 100(4), 241-250. https://doi.org/10.1016/j.anpedi.2024.01.006

Morales, Eva M., Maghioros, María Ángeles, Obregón, Ana M. y Santos, José L. (2020). Adaptación y análisis factorial del cuestionario de conducta de alimentación del adulto (AEBQ) en idioma español. *Revista de la Sociedad Latinoamericana de Nutrición,* 70(1), 40-49. https://doi.org/10.37527/2020.70.1.005

Moreno Redondo, Francisco Javier, Benítez Brito, Néstor, Pinto Robayna, Verta, Ramallo Fariña, Yolanda y Díaz Flores, Carlos (2019). Prevalencia de Trastornos de la Conductra Alimentaria (TCA) en España: necesidad de revisión. *Revista española de nutrición humana y dietética, 23*(1), 130-131. https://doi.org/10.1136/bmj.i5440

Murray, S. B., Nagata, J. M., Griffiths, S., Calzo, J. P., Brown, T. A., Mitchison, D., Blashill, A. J. y Mond, J. M. (2017). The enigma of male eating disorders: A critical review and synthesis. *Clinical Psychology Review, 57*(April), 1-11. https://doi.org/10.1016/j.cpr.2017.08.001

Noriega, David (2023). La anorexia y la bulimia alcanzan ya a las niñas: "Vemos casos con nueve años". elDiario.es. https://www.eldiario.es/sociedad/anorexia-bulimia-ninas_1_10645409.html

O'Brien, Sharon (2023). Human-Centered augmented translation: against antagonistic dualisms. *Perspectives, 32*(3), 391-406. https://doi.org/10.1080/0907676X.2023.224742321

Pérez Fernández, Lucila María yMiaja Menéndez, Patricia (2020). La traducción de los cuestionarios de salud para pacientes. *Káñina, 43*(3), 103-132. https://doi.org/10.15517/rk.v43i3.41935

Quelle, María y García, María (1 de agosto de 2023). Las psicólogas de Lugo avisan: «Hay niñas de ocho años con trastornos alimentarios». *La Voz de Galicia*. https://www.lavozdegalicia.es/noticia/lugo/2023/07/27/psicologas-lugo-avisan-ninas-ocho-anos-trastornos-alimentarios/00031690474413285553924.htm

Rebello, Beatriz Meira, Santos, Giovanna Lima dos, Ávila, Clara Regina Brandão de y Kida, Adriana de Souza Batista (2019). Effects of syntactic simplification on reading comprehension of Elementary School students. *Audiology - Communication Research, 24*(459953), 1-8.

Riley, Anne W. (2004). Evidence that school-age children can self-report on their health. *Ambulatory Pediatrics, 4*(4), 371-376. https://doi.org/10.1367/A03-178R.1

Rivera-Trigueros, Irene (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation, 56*(2), 593-619. https://doi.org/10.1007/s10579-021-09537-5

Salas, Francisca, Hodgson, M. Isabel, Figueroa, Dolly y Urrejola, Pascuala (2011). Características clínicas de adolescentes de sexo masculino con trastornos de la conducta alimentaria. Estudio de casos clínicos. *Revista Medica de Chile, 139*(2), 182-188. https://doi.org/10.4067/S0034-98872011000200007

Savoldi, Beatrice, Gaido, Marco, Bentivogli, Luisa, Negri, Matteo y Turchi, Marco (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics,* 9, 845-874. https://doi.org/10.1162/tacl_a_00401

Schultz, Anna, Maurer, Linda, y Alexandrowicz, Rainer W. (2022). Strengths and weaknesses of the German translation of the Inflexible Eating Questionnaire and of eating disorder assessment in general. *Frontiers in Psychology, 13*(1002463). https://doi.org/10.3389/fpsyg.2022.1002463

Sociedad Española de Médicos Generales y de Familia (2018). Los trastornos de la conducta alimentaria son la tercera enfermedad crónica más frecuente entre adolescentes. www.semg.es

Sousa, Valmi D. y Rojjanasrirat, Wilaiporn (2011). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: A clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice, 17*(2), 268-274. https://doi.org/10.1111/j.1365-2753.2010.01434.x

Anexo 1

ChEAT			
1	I am scared about being overweight.		
Traducción humana			
Me gustaría pesar menos.			
DeepL		**Google Translate**	
Me asusta[4.4] el sobrepeso[2.1].		Me da miedo tener sobrepeso.	
Error type	***Error severity level***	***Error type***	***Error severity level***
Accuracy – Mistranslation[2.1] (cambio de significado) Style – Register[4.4]	Critical (25) Minor (1)	-	-

ChEAT			
2	I stay away from eating when I am hungry.		
Traducción humana			
Intento no comer cuando tengo hambre.			
DeepL		**Google Translate**	
Evito comer cuando tengo hambre.		Evito comer cuando tengo hambre.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	-	-

ChEAT			
3	I think about food a lot of the time.		
Traducción humana			
Pienso mucho en la comida.			
DeepL		**Google Translate**	
Pienso mucho en la comida.		Pienso mucho en la comida.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	-	-

ChEAT			
4	I have gone on eating binges where I feel that I might not be able to stop.		
Traducción humana			
He tenido momentos en los que pensaba que no podía parar de comer.			
DeepL		**Google Translate**	
He tenido atracones en los que creo[3.1;1.3] que no podré[2.1] parar.		He tenido atracones de comida[4.5] en los que siento[3.1] que no podría parar.	
Error type	***Error severity level***	***Error type***	***Error severity level***
Linguistic conventions – Grammar[3.1] (Pto perfecto – Presente) Terminology – Wrong term[1.3] (creer ≠ sentir) Accuracy – Mistranslation[2.1] (el tiempo verbal cambia el sentido de incertidumbre del original)	Minor (1) Major (5) Major (5)	Style – Awkward style[4.5] (redundante) Linguistic conventions – Grammar[3.1] (Pto perfecto – Presente)	Minor (1) Minor (1)
ChEAT			
5	I cut my food into small pieces		
Traducción humana			
Corto la comida en trozos pequeños.			
DeepL		**Google Translate**	
Corto la comida en trozos pequeños.		Corto mi[4.6] comida en trozos pequeños.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	Style – Unidiomatic style[4.6]	Minor (1)

ChEAT			
6	I am aware of the energy (calorie) content in foods that I eat.		
Traducción humana			
Tengo en cuenta la energía que da lo que como.			
DeepL		**Google Translate**	
Soy consciente[2.1] del contenido energético[1.3](calórico)[5] de los alimentos[4.4;1.2] que como [4.5]		Soy consciente[2.1] del contenido energético[1.3](calorías)[5] de los alimentos[4.4;1.2].	
Error type	***Error severity level***	***Error type***	***Error severity level***
Accuracy – Minstranslation[2.1] (ser consciente ≠ considerar/tener en cuenta) Terminology – Wrong term[1.3] (*valor energético*) Audience appropriateness[5] Style – Register[4.4] Terminology – Inconsistent use[1.2] (comida VS alimentos) Style – Awkward style[4.5] (redundante)	Critical (25) Minor (1) Major (5) Minor (1) Minor (1) Minor (1)	Accuracy – Minstranslation[2.1] (ser consciente ≠ considerar/tener en cuenta) Terminology – Wrong term[1.3] (*valor energético*) Audience appropriateness[5] Style – Register[4.4] Terminology – Inconsistent use[1.2] (comida VS alimentos)	Critical (25) Minor (1) Major (5) Minor (1) Minor (1)

ChEAT			
7	I try to stay away from foods such as breads, potatoes, and rice.		
Traducción humana			
Intento no comer comida que me haga subir de peso.			
DeepL		**Google Translate**	
Intento evitar alimentos como el pan, las patatas y el arroz[5].		Intento evitar alimentos como el pan, las patatas y el arroz[5].	
Error type	***Error severity level***	***Error type***	***Error severity level***
Audience appropriateness[5] (¿un niño relaciona esos alimentos con el peso?)	Minor (1)	Audience appropriateness[5] (¿un niño relaciona esos alimentos con el peso?)	Minor (1)

ChEAT			
8	I feel that others would like me to eat more.		
Traducción humana			
Creo que a los demás les gustaría que comiera más.			
DeepL		**Google Translate**	
Siento que a los demás les gustaría que comiera más.		Siento que a los demás les gustaría que comiera más.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	-	-

ChEAT			
9	I vomit after I have eaten.		
Traducción humana			
Vomito después de comer.			
DeepL		**Google Translate**	
Vomito después de comer.		Vomité[3.1] después de comer.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	Linguistic conventions – Grammar[3.1]	Minor (1)

ChEAT			
10	I feel very guilty after eating		
Traducción humana			
Me siento mal (triste) por haber comido.			
DeepL		**Google Translate**	
Me siento muy culpable después de comer.		Me siento muy culpable después de comer.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	-	-

ChEAT			
11	I think a lot about wanting to be thinner.		
Traducción humana			
Pienso mucho en que me gustaría adelgazar.			
DeepL		**Google Translate**	
Pienso mucho en que quiero estar más delgado[2.2;2.1;4.6].		Pienso mucho en querer adelgazar[4.6].	
Error type	***Error severity level***	***Error type***	***Error severity level***
Accuracy – Overtranslation[2.2] (género) Accuracy – Mistranslation[2.1] (estar delgado ≠ adelgazar) Style – Unidiomatic style[4.6]	Minor (1) Major (5) Minor (1)	Style – Unidiomatic style[4.6]	Minor (1)

ChEAT			
12	I think about burning up energy (calories) when I exercise.		
Traducción humana			
Pienso en quemar energía cuando hago deporte.			
DeepL		**Google Translate**	
Pienso en quemar energía[4.6] (calorías[5]) cuando hago ejercicio.		Pienso en quemar energía[4.6] (calorías[5]) cuando hago ejercicio.	
Error type	***Error severity level***	***Error type***	***Error severity level***
Audience appropriateness[5]	Major (5)	Audience appropriateness[5]	Major (5)

ChEAT			
13	Other people think I am too thin.		
Traducción humana			
Creo que los demás piensan que debería pesar más.			
DeepL		**Google Translate**	
Los demás piensan que estoy demasiado delgado[2.2].		Otras personas[4.6] piensan que estoy demasiado delgada[2.2].	
Error type	***Error severity level***	***Error type***	***Error severity level***
Accuracy – Overtranslation[2.2] (género)	Minor (1)	Accuracy – Overtranslation[2.2] (género) Style – Unidiomatic style[4.6]	Minor (1) Minor (1)

ChEAT			
14	I think a lot about having fat on my body.		
Traducción humana			
Pienso en la grasa de mi cuerpo.			
DeepL		**Google Translate**	
Pienso mucho en tener grasa[2.1;4.6] en el cuerpo.		Pienso mucho en tener grasa[2.1;4.6] en el cuerpo.	
Error type	***Error severity level***	***Error type***	***Error severity level***
Accuracy – Mistranslation[2.1] (cambio de significado) Style – Unidiomatic style[4.6] (¿tener grasa en el cuerpo?)	Critical (25) Minor (1)	Accuracy – Mistranslation[2.1] (cambio de significado) Style – Unidiomatic style[4.6] (¿tener grasa en el cuerpo?)	Critical (25) Minor (1)

ChEAT			
15	I take longer than others to eat my meals.		
Traducción humana			
Tardo más que los demás en comer.			
DeepL		**Google Translate**	
Tardo más que los demás en comer.		Tardo más que los demás en comer.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	-	-

ChEAT			
16	I stay away from foods with sugar in them.		
Traducción humana			
Intento no comer cosas con azúcar.			
DeepL		**Google Translate**	
Evito los alimentos[1.2;4.4] con azúcar.		Evito los alimentos[1.2;4.4] con azúcar.	
Error type	***Error severity level***	***Error type***	***Error severity level***
Style – Register[4.4] Terminology – Inconsistent use[1.2] (comida VS alimentos)	Minor (1) Minor (1)	Style – Register[4.4] Terminology – Inconsistent use[1.2] (comida VS alimentos)	Minor (1) Minor (1)

ChEAT			
17	I eat diet foods.		
Traducción humana			
Tomo comidas ligeras.			
DeepL		**Google Translate**	
Como alimentos dietéticos[5].		Como alimentos dietéticos[5].	
Error type	***Error severity level***	***Error type***	***Error severity level***
Audience appropriateness[5] Terminology – Inconsistent use[1.2] (comida VS alimentos)	Major (5) Minor (1)	Audience appropriateness[5] Terminology – Inconsistent use[1.2] (comida VS alimentos)	Major (5) Minor (1)

ChEAT			
18	I think that food controls my life.		
Traducción humana			
Creo que la comida controla mi vida.			
DeepL		**Google Translate**	
Creo que la comida[1.2] controla mi vida.		Creo que la comida[1.2] controla mi vida.	
Error type	***Error severity level***	***Error type***	***Error severity level***
Terminology – Inconsistent use[1.2] (comida VS alimentos)	Minor (1)	Terminology – Inconsistent use[1.2] (comida VS alimentos)	Minor (1)

ChEAT			
19	I can show self-control around food.		
Traducción humana			
Me sé controlar al comer.			
DeepL		**Google Translate**	
Puedo mostrar autocontrol con respecto a la comida.		Me controlo[2.1] con respecto a la comida.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	Accuracy – Mistranslation[2.1] (cambio de significado): controlarse -> restricción ≠autocontrol	Critical (25)

ChEAT			
20	I feel that others pressure me to eat.		
Traducción humana			
Creo que los demás me presionan para que coma más.			
DeepL		**Google Translate**	
Siento que los demás me presionan para que coma.		Siento que los demás me presionan para comer[1.1].	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	Accuracy – Undertranslation[1.1] (para comer ¿más deprisa?) Demasiado general, no enfatiza que la presión está dirigida hacia la persona, sino hacia el acto de comer	Major (5)

ChEAT			
21	I give too much time and thought to food.		
Traducción humana			
Pienso mucho tiempo en la comida.			
DeepL		**Google Translate**	
Dedico demasiado tiempo y reflexión[4.5;4.4] a la comida.		Dedico demasiado tiempo y pensamiento[.5;4.4] a la comida.	
Error type	***Error severity level***	***Error type***	***Error severity level***
Style – Awkward style[4.5] (redundante) Style – Register[4.4] (¿reflexión?)	Minor (1) Minor (1)	Style – Awkward style[4.5] (redundante) Style – Register[4.4]	Minor (1) Minor (1)

ChEAT			
22	I feel uncomfortable after eating sweets.		
Traducción humana			
Me siento mal (triste) después de comer dulces.			
DeepL		**Google Translate**	
Me siento incómodo[2.2;4.6] después de comer dulces.		Me siento incómodo[2.2;4.6] después de comer dulces.	
Error type	***Error severity level***	***Error type***	***Error severity level***
Accuracy – Overtranslation[2.2] (género) Style – Unidiomatic style[4.6] ¿sentirse incómodo?	Minor (1) Minor (1)	Accuracy – Overtranslation[2.2] (género) Style – Unidiomatic style[4.6] ¿sentirse incómodo?	Minor (1) Minor (1)

ChEAT			
23	I have been dieting.		
Traducción humana			
He comido para no engordar.			
DeepL		**Google Translate**	
He estado a dieta[5].		He estado a dieta[5].	
Error type	***Error severity level***	***Error type***	***Error severity level***
Audience appropriateness[5]	Major (5)	Audience appropriateness[5]	Major (5)

ChEAT			
24	I like my stomach to be empty.		
Traducción humana			
Me gusta tener la barriga vacía.			
DeepL		**Google Translate**	
Me gusta tener el estómago vacío.		Me gusta tener el estómago vacío.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	-	-

ChEAT			
25	I enjoy trying new rich foods.		
Traducción humana			
Me gusta probar platos nuevos y nutritivos			
DeepL		**Google Translate**	
Me gusta probar nuevos alimentos ricos[1.3;1.2].		Disfruto[4.4] probando nuevos alimentos ricos[1.3;1.2].	
Error type	***Error severity level***	***Error type***	***Error severity level***
Terminology – Wrong term[1.3] (alimentos ricos ≠ ricos en nutrientes/calorías/abundantes) Terminology – Inconsistent use[1.2] (comida VS alimentos)	Critical (25) Minor (1)	Terminology – Wrong term[1.3] (alimentos ricos ≠ ricos en nutrientes/calorías/abundantes) Terminology – Inconsistent use[1.2] (comida VS alimentos) Style – Register[4.4] (¿disfruto?)	Critical (25) Minor (1) Minor (1)

ChEAT			
26	I have the urge to vomit after eating.		
Traducción humana			
Tengo ganas de vomitar después de comer.			
DeepL		**Google Translate**	
Tengo ganas de vomitar después de comer.		Tengo ganas de vomitar después de comer.	
Error type	***Error severity level***	***Error type***	***Error severity level***
-	-	-	-

Absolute Penalty Total (APT)	155	Absolute Penalty Total (APT)	147
Per-Word Penalty Total (PWPT)	0,8424	Per-Word Penalty Total (PWPT)	0,8167
Overall Normed Penalty Total (ONPT)	181,11	Overall Normed Penalty Total (ONPT)	175,58
Overall Quality Score (OQS)	15,76	Overall Quality Score (OQS)	18,33
Overall Quality Fraction (OQF)	0,16	Overall Quality Fraction (OQF)	0,18

Evaluation Word Count (EWC):	184
Reference Word Count (RWC):	215
Penalty Scaler (PS):	1,00
Max. Score Value (MSV):	100,00

Evaluation Word Count (EWC):	180
Reference Word Count (RWC):	215
Penalty Scaler (PS):	1,00
Max. Score Value (MSV):	100,00

MQM Scorecard: Full MQM-Core Error Typology with 4 Severity Levels (https://themqm.org/wp-content/uploads/2022/08/Scoring-Model-Calculator-2022-08-22_SEW_Distribution.xlsx)

Error type

- *Terminology*[1]
  - Inconsistent with terminology resource[1.1]
  - Inconsistent use of terminology[1.2]
  - Wrong term[1.3]
- *Accuracy*[2]
  - Mistranslation[2.1]
  - Overtranslation[2.2]
  - Undertranslation[2.3]
  - Addition[2.4]
  - Omission[2.5]
  - Do not translate[2.6]
  - Untranslated[2.7]
- *Linguistic conventions*[3]
  - Grammar[3.1]
  - Punctuation[3.2]
  - Spelling[3.3]
  - Unintelligible[3.4]
  - Character encoding[3.5]
  - Textual conventions[3.6]

- Style[4]
    - Organization style[4.1]
    - Third-party style[4.2]
    - Inconsistent with external reference[4.3]
    - Language register[4.4]
    - Awkward style[4.5]
    - Unidiomatic style[4.6]
    - Inconsistent style[4.7]
- Audience appropriateness[5]
    - Culture-specific reference[5.1]
    - Offensive[5.2]

Error Severity Levels:	**Neutral**	**Minor**	**Major**	**Critical**
Severity Penalty Multipliers:	0	1	5	25

MQM Scorecard: Full MQM-Core Error Typology with 4 Severity Levels (https://themqm.org/wp-content/uploads/2022/08/Scoring-Model-Calculator-2022-08-22_SEW_Distribution.xlsx)

- **Neutral (0 puntos):** No afecta la comprensión ni la calidad de la traducción.
- **Minor (1 punto):** Errores leves de estilo, fluidez o gramática sin impacto en el significado.
- **Major (5 puntos):** Errores graves que dificultan la comprensión o distorsionan parcialmente el significado.
- **Critical (25 puntos):** Errores críticos que cambian completamente el significado o generan información incorrecta.

Calculation:
ONPT=PWPT × RWC x PS
OQF=1 – (ONPT / RWC)
OQS = OQF × MSV
OQS=1-(APT/EWC)