Traducción automática neuronal y traducción automática estadística: percepción y productividad

Ariana López Pereira

revista.tradumàtica
tecnologies de la traducció

Ariana López Pereira
Grup Tradumàtica, Universitat Autònoma de Barcelona
ariana.lopezp@e-campus.uab.cat;
ORCID: 0000-0003-4728-2108

Resumen

El campo de la traducción automática ha cambiado por completo con los avances que han experimentado los motores de traducción automática neuronal (TAN), especialmente en comparación con los resultados que se estaban obteniendo con los de traducción automática estadística (TAE). Así, se hace necesario revisar su uso y percepción por parte de los usuarios finales: los traductores. El objetivo principal de este trabajo es determinar la percepción y la productividad, en términos de tiempo y número de modificaciones, de un grupo de traductores al utilizar sistemas de TAE y de TAN. Para ello, mediante la plataforma Dynamic Quality Framework (DQF) de TAUS, diez traductores profesionales evaluaron en primer lugar los segmentos de traducción automática en bruto de dos textos —un manual de instrucciones y un texto de marketing—, propuestos por el motor de Microsoft Translation (TAE) y de Google Neural Machine Translation (TAN). Posteriormente, seis de los diez traductores poseditaron dos pruebas de productividad para determinar el tiempo y la distancia de edición. Los resultados mostraron que, en opinión de los traductores, sería más productivo el motor neuronal, ya que, según su percepción, la posedición de sus propuestas conllevaría menos tiempo y menos correcciones. Sin embargo, al cotejar estos resultados con los obtenidos en las pruebas de productividad, se observó que, aunque la distancia de edición era menor con el motor de TAN que con el de TAE, el tiempo de posedición era mucho mayor para el motor neuronal.

Palabras clave: traducción automática neuronal, traducción automática estadística, distancia de edición, productividad, percepción, posedición

Resum

El camp de la traducció automàtica ha canviat per complet amb els progressos que han experimentat els motors de traducció automàtica neuronal (TAN), sobretot si es compara amb els resultats obtinguts amb els de traducció automàtica estadística (TAE). Així, és necessari revisar-ne l'ús i la percepció per part dels usuaris finals, els traductors. L'objectiu principal d'aquest treball és determinar la percepció i la productivitat, en termes de temps i nombre d'edicions, d'un grup de traductors a l'hora d'utilitzar sistemes de TAE i de TAN. Amb aquest objectiu, mitjançant la plataforma Dynamic Quality Framework (DQF) de TAUS, deu traductors professionals han avaluat, primerament, els segments de traducció automàtica en brut de dos textos, un

Rebuda: 16 de novembre de 2019 | Acceptació: 9 de desembre de 2019 | Publicació avançada: 23 de desembre de 2019

manual d'instruccions i un text de màrqueting, proposats pel motor de Microsoft Translation (TAE) i de Google Neural Machine Translation (TAN). Posteriorment, sis dels deu traductors han posteditat dues proves de productivitat, a fi d'establir-ne el temps i la distància d'edició. Els resultats mostren que els traductors consideren el motor neuronal més productiu, atès que, segons la seva percepció, triguen menys temps a posteditar, la qual cosa comporta menys edicions. No obstant això, en comparar aquests resultats amb els obtinguts a les proves de productivitat, encara que la distància d'edició és menor amb el motor de TAE que amb el de TAN, el temps de postedició és molt més alt en el cas del motor neuronal.

Paraules clau: traducció automàtica neuronal, traducció automàtica estadística, distància d'edició, productivitat, percepció, postedició

Abstract

The machine translation field has changed completely due to the many advances seen in neural machine translation (NMT) engines, especially in comparison with the results that were obtained with statistical machine translation (SMT) engines. So, it is necessary to review not only how MT is used but also how it is perceived by the end users, the translators. The main objective of this study is to determine the perception and productivity of a group of translators using SMT and NMT systems in terms of time and edit distance. Via the TAUS Dynamic Quality platform, ten professional translators first evaluated raw machine translation segments from two different texts – a user guide and a marketing text – proposed by the Microsoft Translation Engine (SMT) and Google Neural Machine Translation (NMT). Six of the ten translators subsequently post-edited two productivity tests to determine time and edit distance. The results show that translators perceive the NMT system as more productive because, according to their perception, it would take less time to post-edit and would mean fewer editions. However, when comparing these results with those obtained in productivity tests, although the edit distance was shorter when using the SMT engine than with the NTM, the post-editing time is much longer for the neural engine.

Keywords: Neural Machine Translation, Statistical Machine Translation, productivity, edit distance, post-editing time, perception, post-editing

Introducción y literatura relacionada

En los últimos años, el panorama de la traducción y, concretamente, de la traducción automática ha cambiado radicalmente, debido a la irrupción en el mercado de la traducción automática neuronal (TAN). Los grandes avances en tecnología y en el campo de la inteligencia artificial han propiciado que, en la actualidad, los sistemas de traducción automática cambien y se desarrollen rápidamente. Por su parte, las necesidades del sector siguen evolucionando: se está volviendo imprescindible traducir un número cada vez mayor de palabras en el menor tiempo posible, así como abaratar los costes. Ante este panorama, el empleo de la traducción automática se está convirtiendo en algo tan esencial como habitual. Sin embargo, esto debe darse siempre y cuando se pueda establecer un estándar de calidad y se vele por que el coste que implique, económico o de esfuerzo, se vea compensado por las ventajas y los beneficios.

Motivadas precisamente por esta reducción de los costes y en aras del aumento de la productividad, cada vez son más las empresas que han incorporado la traducción automática a su flujo de trabajo, como se puede ver, por ejemplo, en el informe realizado por Torres-Hostench, Presas y Cid-Leal (2016). No obstante, puesto que la traducción

automática en bruto en muchas ocasiones no está exenta de errores, es necesario que los resultados propuestos se validen o modifiquen para garantizar su calidad. Así, la traducción automática se incorpora al flujo de trabajo mediante la posedición, entendida como la edición y la corrección de las propuestas de traducción automática (ISO, 2017). Asimismo, dado que la posedición se puede aplicar en distintos grados, es posible adaptarla con facilidad a las distintas necesidades del mercado (Way, 2013; Aramberri, 2014). Por consiguiente, la traducción automática facilita el incremento de la productividad, permite traducir contenido que, de otra forma, no sería posible traducir y reduce los tiempos y los costes en el mercado (Way, 2013).

A pesar de todo esto, en ocasiones parece que los avances señalados no son propicios para el usuario final, ya que las tarifas son cada vez más bajas y los traductores consideran que la traducción automática supone una amenaza para su trabajo, pues les perjudica a nivel creativo y profesional (Guerberof, 2013; Moorkens, 2017). Además, este no es el único desafío al que se enfrentan los traductores profesionales: en la actualidad, las *crowd translations* (traducciones realizadas por no profesionales) son cada vez más comunes (Katan, 2016).

Ante este panorama tan cambiante, el presente estudio se propone abordar cómo valoran los traductores profesionales los resultados en bruto de la traducción automática mediante dos sistemas de traducción, uno neuronal y uno estadístico, y correlacionar los datos obtenidos con la productividad real a la hora de poseditar textos. A tal fin, este artículo se refiere, en primer lugar, a una serie de estudios previos relacionados con la investigación que aquí se presenta, en los cuales se abordan tanto los aspectos comparativos como el contexto en el que estos se dan (desde el punto de vista de la comparación y del uso de ambos sistemas, así como de la percepción de los usuarios al emplearlos). En segundo lugar, se incluye el apartado de los objetivos y, posteriormente, el de la metodología, en el que se describen los instrumentos y las pruebas realizadas para recopilar todos los datos. Seguidamente, se exponen los resultados y, por último, se presentan las conclusiones y se proporcionan algunas líneas para continuar la investigación.

Los estudios que se presentan a continuación permiten ilustrar el contexto que acabamos de describir. Algunos de ellos son relevantes por la relación que guardan con la metodología de este artículo (entre ellos, la investigación de Guerberof del año 2009, en la que se realizan pruebas de posedición tanto con un motor de traducción estadístico como con memorias de traducción, y el estudio de Bentivogli *et al.* del año 2016, una investigación comparativa que analiza segmentos poseditados de un motor de traducción automática neuronal y de un motor de traducción automática estadística), mientras que otros destacan por su relación con los objetivos presentados aquí (por ejemplo, Guerberof, 2013). También cabe resaltar, debido a su metodología, la investigación de Shterionov *et al.* (2017). Sin embargo, hay un nicho de investigación que queda por abordar y al que intentaremos dar respuesta aquí: la relación entre la percepción de los traductores, usuarios finales de la posedición, y los resultados que se obtienen al poseditar con un motor de traducción automática neuronal y un motor de traducción automática estadística.

El estudio de Guerberof (2009) aborda algunos de los aspectos que hemos mencionado en la introducción. En él, se realizó una valoración con nueve traductores profesionales a través de una herramienta de posedición en línea con la cual debían traducir segmentos nuevos y poseditar segmentos de traducción automática estadística y de una memoria de traducción (coincidencias parciales entre el 80 % y el 90 %) sin conocer su procedencia. Los resultados sobre la productividad indicaron que los traductores eran más rápidos poseditando que traduciendo los segmentos de cero y que esto también sucedía cuando editaban las coincidencias parciales de la memoria de traducción. En cuanto a los errores encontrados al evaluar la calidad, más de la mitad procedían de los segmentos de la memoria de traducción, mientras que el 27 % se habían realizado en los segmentos de traducción automática y el 21 %, en los segmentos nuevos. Con respecto a la clasificación de estos errores, el 44 % (casi la mitad) eran de precisión, mientras que el resto se dividía en errores de idioma (26 %), de traducción (16 %), de terminología (16 %) y de consistencia (solo un 1 %). No resulta posible conocer el porcentaje de errores de fluidez, ya que, para este trabajo, se utilizó la escala de LISA, en la que la fluidez no figura como categoría.

La investigación de Bentivogli *et al.* (2016), por su parte, es uno de los primeros estudios comparativos de los resultados que se pueden obtener al analizar los segmentos propuestos por un motor de traducción automática neuronal y uno de traducción automática estadística basado en frases. Uno de los objetivos de esta investigación era determinar en qué contexto un sistema neuronal presentaba mejores resultados que un sistema estadístico. Para ello, se emplearon datos de evaluación del IWSLT del año 2015 de traducción automática en la combinación de idiomas inglés-alemán, lo que permitía disponer no solo de tres sistemas estadísticos de última generación, sino también de posediciones realizadas por traductores profesionales. Otro de los motivos para emplear esta combinación de idiomas era el hecho de que se trataba de un par lingüístico con diferencias morfológicas y de orden de palabras muy patentes.

En esta investigación, se analizaron errores morfológicos, léxicos y de ordenación de palabras en doce charlas TED, lo que supuso un total de 600 frases y aproximadamente 10 000 palabras. Los resultados mostraron que el motor de TAN era considerablemente superior a los de TAE, puesto que el esfuerzo de posedición se vio reducido, de media, en un 26 %. Además, el sistema de TAN era superior en evaluación al TAE, independientemente de la longitud de las frases (aunque empeoraba rápidamente en comparación con el resto de sistemas a medida que las frases eran más largas). Por otra parte, el motor neuronal mostraba ventajas a la hora de poseditar textos léxicamente ricos, y sus resultados contenían menos errores léxicos (-19 %), menos errores de ordenación de palabras (-50 %) y una cantidad considerablemente menor de errores en la colocación de los verbos (-70 %).

Otro estudio comparativo es el de Esperança-Rodier *et al.* (2017), que comparte características metodológicas con el presentado en este artículo, como se verá más adelante. Su investigación se centra en el efecto que tienen los sistemas de traducción automática neuronal y estadística en la actividad de los traductores, mediante la identificación de los usos y las percepciones de los estudiantes de un máster con

sistemas de TAN y TAE, y compara estos usos y percepciones con la arquitectura y funcionalidad de los sistemas. Para llevar a cabo este estudio se empleó, primero, el corpus de la campaña de evaluación de IWSLT del año 2010 en la combinación francés-inglés, así como dos motores propios. Se realizó una evaluación empírica de la calidad en términos de fluidez y precisión de los dos sistemas de traducción automática, al igual que en este trabajo. Por otra parte, se otorgó una determinada puntuación según la traducción fuera —en términos de los propios autores— mala (cuando no era fluida y/o precisa), de calidad intermedia (cuando era adecuada y/o precisa) o buena (cuando era adecuada y precisa). Posteriormente, se escogieron 50 frases para evaluarlas, en este caso, en base a cuatro categorías, que abarcaban desde una traducción muy mala hasta una muy buena (de nuevo, según la terminología de los autores). Esta es una categorización similar a la empleada en las pruebas de evaluación de calidad de los segmentos de TAN y TAE que se tratarán más adelante en este artículo.

En segundo lugar, para determinar la percepción de los estudiantes, se crearon dos tareas: la primera consistía en corregir los resultados de traducción automática del motor neuronal de Google y del motor MT@EC, el motor estadístico de la Comisión Europea, mientras que la segunda radicaba en alternar trabajos de traducción y posedición. Seguidamente, los estudiantes debían responder a un cuestionario de veinte preguntas. Se llevó a cabo una comparación cuantitativa a través de las métricas automáticas BLEU (Papineni *et al.*, 2002), TER (Translation Edit Rate) (Snover *et al.* 2006) y METEOR (Wołk y Koržinek, 2017). Además, se realizó un análisis cualitativo de los errores de traducción según criterios lingüísticos (Vilar, 2006), para determinar qué patrones sintácticos implicaban errores de traducción y qué tipo de error era el más común. En último lugar, se llevó a cabo el análisis estadístico. En este caso, a diferencia del estudio de Bentivogli *et al.* (2016), los resultados de los dos motores fueron equivalentes, aunque el uso del motor neuronal ofrecía unos resultados ligeramente mejores en la puntuación de las métricas automáticas. Por otra parte, la evaluación humana presentaba una correlación con los resultados obtenidos a partir de las métricas automáticas. Con respecto a la percepción, los estudiantes consideraron negativamente los dos motores por igual.

La evaluación de motores neuronales y estadísticos no se ha efectuado únicamente en el ámbito académico. Dada su importancia, la industria también ha jugado un papel fundamental. Una de las evaluaciones de estos sistemas llevada a cabo en el contexto empresarial es la de Kantan MT (Shterionov *et al.*, 2017). En este trabajo se compara la calidad de un motor estadístico basado en frases y de un motor neuronal en una plataforma personalizada (Custom Machine Translation de Kantan) para la producción de traducción a gran escala. Para este estudio se crearon cinco motores de TAN y cinco motores de TAE en las combinaciones del inglés al alemán, chino, japonés, italiano y español con el mismo grupo de datos y siguiendo exactamente el mismo procedimiento. Para la evaluación se emplearon las métricas automáticas F-measure, BLEU y TER, así como una evaluación humana realizada por quince estudiantes del Máster de Traducción Especializada Multilingüe (Multilingual Specialized Translation), tres por cada lengua, todos ellos hablantes nativos. En esta prueba, los revisores empleaban la herramienta de evaluación en línea propia de la empresa y debían comparar el segmento de origen con

las dos opciones propuestas por los dos sistemas y decidir cuál de las dos era de mejor calidad o si presentaban la misma calidad. Los resultados obtenidos muestran que, en todos los casos, el motor de TAN recibió mejor valoración que el de TAE basado en frases. Sin embargo, estos datos no se corresponden exactamente con los de BLEU, pues esta métrica parece subestimar la calidad del motor neuronal.

También desde el punto de vista del uso se han realizado estudios de comparación entre un motor neuronal y un motor estadístico. En este caso, la evaluación se centra en contenido generado por el usuario; más concretamente, en tuits en inglés y alemán (Lohar *et al.*, 2019). Así, se crearon dos sistemas: uno de traducción automática neuronal y otro estadístico basado en frases. El objetivo era comparar la traducción automática de los tuits a través de estos dos sistemas y evaluar las diferencias en términos de cantidad y tipos de corpus a la hora de entrenar cada uno de los motores.

Se tradujeron tuits del alemán al inglés con un motor de TAN y uno de TAE entrenados con el conjunto específico FIFA 2014, con únicamente 4000 pares de tuits, y, posteriormente, con el conjunto de tuits de Harvard no específico de inglés y con un corpus de segmentos cortos de comentarios de noticias inglés-alemán, ya que los tuits son, por definición, textos cortos. Puesto que el conjunto de Harvard solo contenía tuits en inglés, fue necesario poseditarlo primero al alemán para incluirlos. Para el motor estadístico se empleó Moses, mientras que se utilizó OpenNMT como motor de traducción automática neuronal. Para la evaluación de los sistemas, también en este caso se emplearon las métricas BLEU, TER y METEOR. Los resultados mostraron que el desempeño del motor neuronal era considerablemente menor cuando el conjunto de datos era pequeño (en este caso, únicamente el de FIFA 2014). No obstante, cuando los motores fueron alimentados con la selección de Harvard y, posteriormente, con el corpus de noticias, los resultados del motor de TAN mejoraron en gran medida, mientras que los del motor estadístico lo hicieron solo ligeramente.

En relación con el uso de la traducción automática, cabe destacar que se han realizado investigaciones en campos muy variados, pero incluimos el siguiente ejemplo porque también es un estudio comparativo de la posedición de los segmentos propuestos por un motor de TAN y por uno de TAE. En este caso, Toral *et al.* (2018) llevaron a cabo una investigación de carácter comparativo en el campo de la traducción literaria. En esta investigación, seis traductores profesionales con experiencia en este campo trabajaron en la producción de traducciones literarias del inglés al catalán utilizando tres procedimientos diferentes: mediante la traducción desde cero, por medio de la posedición de los resultados de un motor estadístico y mediante la posedición de los resultados de un motor neuronal. Ambos motores eran de dominio específico y habían sido entrenados con 133 novelas, lo que equivalía a un millón de palabras. Para realizar las pruebas de traducción, se escogió una novela publicada con una licencia Creative Commons, a fin de garantizar la replicabilidad del estudio. Se realizaron entrevistas y cuestionarios para conocer las opiniones y percepciones de los traductores. Con respecto al esfuerzo temporal, se observó que todos los traductores trabajaban más rápido al poseditar los segmentos de TAE que al traducir de cero. Al mismo tiempo, se vio que los sujetos poseditaban a mayor velocidad cuando trabajaban sobre los segmentos de

TAN que cuando lo hacían partiendo de los de TAE. Sin embargo, la percepción de los traductores sobre la TAE no se correspondía con estos resultados: dos de ellos, por ejemplo, consideraban que eran más lentos poseditando estos segmentos que traduciendo de cero. Siguiendo con los resultados de percepción, los traductores consideraban que, al poseditar, eran menos creativos, ya que se veían condicionados por la propuesta. Asimismo, observaron, los sistemas calcaban demasiado la estructura original del inglés.

Por su parte, Guerberof (2013) también investigó el concepto de percepción en la posedición llevada a cabo por traductores profesionales, al igual que en el presente artículo. Para ello, reunió a un grupo de 24 traductores y tres revisores, quienes debían completar un cuestionario para reflejar su opinión sobre este proceso. Todos los traductores tenían una dilatada experiencia en el campo de la localización, de entre dos y ocho años. Los resultados del cuestionario, recopilados a través de la plataforma SurveyMonkey, señalaban que, según un 45 % de los participantes, al emplear posedición, su productividad se mantenía constante con el paso del tiempo. Sin embargo, según un 40 %, su productividad aumentaba con el paso del tiempo. Por otra parte, se comprobó que, conforme aumentaba su experiencia, para el 55 % era más fácil detectar los errores de la traducción automática, mientras que el 30 % consideraba que su experiencia no se veía afectada. En ambos casos, el 15 % de los participantes no sabía qué responder. Al preguntarles sobre el esfuerzo de posedición, el 30 % consideraba que poseditar requería el mismo esfuerzo que revisar traducciones humanas; el 40 % creía que necesitaba más esfuerzo al poseditar que al revisar estas traducciones y solo el 20 % postulaba que era necesario menos esfuerzo al poseditar que al revisar traducciones humanas. Un 10 % no daba a conocer su opinión.

De los trabajos mencionados con anterioridad, cabría destacar especialmente el de Bentivogli *et al.* (2016), debido a su relevancia como primer estudio comparativo de los sistemas de TAN y TAE, así como por su importancia en cuanto a la metodología para este artículo, dadas las similitudes que existen entre ambos. Asimismo, es esencial resaltar los objetivos que se presentan en el estudio de Guerberof (2009 y 2013) y la metodología empleada en el artículo de Shterionov *et al.* (2017), en el que también se realiza una comparación entre resultados de un motor de TAN y uno de TAE, si bien en su caso se lleva a cabo a gran escala. De esta forma, con esta investigación se pretende analizar las percepciones de los poseditores y cotejarlas con los resultados de sus propias posediciones, un nicho de investigación que no se había abordado con anterioridad.

Objetivos e hipótesis

En primer lugar, el objetivo de este estudio es determinar la percepción de los traductores a la hora de valorar segmentos de traducción automática neuronal y traducción automática estadística, a fin de averiguar si, en su opinión, los segmentos de traducción automática en bruto propuestos por el sistema neuronal son más productivos; esto es: si son más rápidos de poseditar y requieren un menor número de ediciones. Para explorar las percepciones de los traductores, se usará una prueba de

MT Ranking del DQF de TAUS. Por otro lado, se parte de la hipótesis de que los traductores valorarán mejor la TAN, pues se prevé que la poseditarán en menos tiempo y requerirá menos cambios. Para llevar a cabo el estudio, se emplearán dos textos: uno de marketing y otro extraído de un manual de instrucciones.

En segundo lugar, y mediante los mismos segmentos empleados para la evaluación de la percepción, se analizará la productividad a través de la posedición de estos. El objetivo es descubrir si la productividad, en términos de tiempo y distancia de edición, es mayor al utilizar un motor u otro. Para ello, se comprobará cuánto tiempo se tarda en poseditar tanto el texto de marketing como el del manual de instrucciones con el motor de TAN y con el de TAE. En este caso, se parte de la hipótesis de que existen diferencias en cuanto a esfuerzo temporal y técnico (Krings, 2001) en los resultados obtenidos a partir de la posedición de los segmentos de traducción automática neuronal y de la traducción automática estadística.

Para obtener los resultados sobre la percepción de los traductores, se realizará una prueba dentro del Dynamic Quality Framework de la herramienta TAUS. De la misma manera, para determinar la productividad, se utilizará una prueba de productividad dentro de este mismo marco. Se emplearán dos tipos de textos, extraídos de un manual de instrucciones y de productos de marketing, y los motores de traducción automáticos serán Google Neural Machine Translation, para el motor neuronal, y Microsoft Translator, para el motor de estadística. En el momento en el que se realizó este proceso (marzo de 2018), Microsoft todavía no había lanzado su motor neuronal y los resultados proporcionados eran únicamente estadísticos. Diez participantes efectuarán las pruebas de evaluación de cada motor (MT Ranking) y, de estos diez, seis trabajarán en las pruebas de posedición, que permitirán determinar la productividad.

Metodología

A continuación, se describe la metodología empleada en esta investigación. En primer lugar, se aborda la descripción de las pruebas empleadas para alcanzar los objetivos presentados con anterioridad. Posteriormente, se describe la preparación de las pruebas y, por último, se ofrece un análisis de los resultados obtenidos.

Descripción de las pruebas

Para llevar a cabo este estudio, se recurrió a distintas pruebas adaptadas a los objetivos que se querían conseguir. Todas ellas se realizaron en el Dynamic Quality Framework de TAUS. Así, para obtener los resultados de percepción de los diez participantes sobre la calidad de los motores de TAN y TAE, se empleó el MT Ranking, que TAUS define como una tarea de comparación «que ayuda a los usuarios a seleccionar motores de traducción automática o traductores humanos en función de la calidad del resultado» (Görög, 2017, traducción propia). Por su parte, para determinar el tiempo empleado por los sujetos a la hora de poseditar los distintos textos, se empleó una prueba de

productividad. Esta se define de la siguiente manera (Görög, 2014, 452, traducción propia):

> Los usuarios pueden optar entre poseditar todos los segmentos de traducción automática o traducir la mitad de los segmentos de cero y poseditar la otra mitad. En este último caso, la herramienta de DQF elimina la mitad de los segmentos (de traducción automática) del archivo o archivos subidos. En ambos casos, el sistema mide la distancia de edición y el tiempo empleado para completar las tareas. Al asignar la tarea a los usuarios, es necesario especificar cuál de los dos tipos de posedición se requiere (posedición completa o posedición parcial).

Para esta investigación, se escogió poseditar todos los segmentos con la instrucción de *full post-editing* o posedición completa; esto es: se debían realizar tantas correcciones como fueran necesarias para alcanzar el estándar de calidad humano. En el caso de la *light post-editing* o posedición parcial, por otra parte, el número de modificaciones es el mínimo necesario para obtener un texto comprensible.

Además de tener en cuenta la percepción de los traductores, y con el objetivo de triangular mejor los resultados obtenidos en la prueba de productividad, se llevó a cabo una prueba de evaluación de calidad de los dos motores, bajo los criterios de fluidez y precisión. TAUS adopta para estos valores las definiciones del Linguistic Data Consortium. Así, la precisión viene definida como «el grado en el que el significado expresado en la traducción de referencia o el texto original se expresa en la traducción de destino», mientras que la fluidez es «el grado en el que la traducción es correcta gramaticalmente, presenta una ortografía correcta, se adhiere al uso común de términos, títulos y nombres, es aceptable de forma intuitiva y puede ser interpretada con sensatez por un hablante nativo» (Görog, 2014, traducción propia). Para esta evaluación se emplearon cuatro pruebas distintas, a fin de abarcar todas las combinaciones entre documentos y motores.

Las pruebas de MT Ranking y de evaluación de calidad se realizaron simultáneamente. En el caso de la calidad, una única persona efectuó cuatro pruebas que servían para determinar la fluidez y precisión de los mismos segmentos con los que los participantes estaban trabajando en la prueba de MT Ranking. Las cuatro evaluaciones se correspondían con las cuatro variaciones posibles: la evaluación de los segmentos propuestos tanto por el motor estadístico como por el neuronal para el texto de marketing y el del manual de usuario. La prueba consistía en que el sujeto determinase la fluidez y la precisión de forma simultánea en cuatro niveles. En términos de fluidez, los segmentos recibían las siguientes puntuaciones: Incomprensible (1), Poco fluido (2), Bueno (3) y Sin errores (4). La precisión también se evaluaba en cuatro grados: Nada (1), Poca (2), Bastante (3) y Toda (4). Si bien los resultados obtenidos aquí responden a un criterio cualitativo, aunque con datos cuantitativos, estos se pueden cotejar posteriormente con el número de palabras por segmento, para analizar mejor los resultados obtenidos.

Preparación de las pruebas

Previamente a efectuar las pruebas, fue necesario preparar los distintos instrumentos para realizarlas. En primer lugar, se escogieron los textos sobre los que se efectuarían la evaluación de calidad y el test de productividad. Para ello se eligieron dos fragmentos de textos ya publicados en línea: por un lado, se trabajó con un texto de un manual de usuario de un *smartwatch* y, por otro, con un texto de marketing de un *smartphone*. Si bien los textos tenían una longitud total considerable, se los acortó hasta obtener aproximadamente 2000 palabras en cada uno. Tras esto, se preparó un proyecto en la herramienta TAO SDL Trados Studio (versión 2017) para crear las distintas memorias de traducción, en formato TMX, que se añadirían posteriormente a DQF.

Simultáneamente, se crearon las dos API de los dos motores de traducción automática. A través de la plataforma Google Cloud, se creó la API para el motor de TAN de Google. Por su parte, se empleó la plataforma de Microsoft Azure para crear la API de conexión al sistema TAE. Como se ha mencionado con anterioridad, cuando se creó la API para el motor de TAE, Microsoft todavía no había dado acceso al público a su motor neuronal y los resultados proporcionados eran únicamente estadísticos. Se añadieron las API a la configuración del proyecto de SDL Trados Studio (en el caso de Microsoft, mediante el *plug-in* MT Enhanced) y, mediante la tarea Pretraducir, se descargaron todos los segmentos de traducción automática en bruto. Se volcaron los resultados en las memorias de traducción, lo que dio lugar a cuatro memorias que se correspondían con cada texto y tipo de motor, a saber: Marketing – TAE, Marketing – TAN, Documentación – TAE y Documentación – TAN, y, a continuación, se las exportó al formato TMX, aceptado por la plataforma. Cada una de ellas contaba con 250 segmentos, el mínimo válido en la herramienta DQF. Para la prueba de MT Ranking fue necesario juntar las cuatro memorias en una.

En cuanto a los sujetos, un total de diez traductores participaron en la primera prueba realizada, la de MT Ranking. Todos ellos tenían edades comprendidas entre los 25 y los 32 años y habían cursado el Grado de Traducción e Interpretación o la Licenciatura en Traducción e Interpretación. Su experiencia profesional era de entre uno y tres años en el campo de la traducción, ya fuera como traductores autónomos o como traductores o gestores de proyectos en empresas proveedoras de servicios de idiomas (LSP, por sus siglas en inglés) o *multilanguage vendors* (MLV), y en aquel momento se encontraban trabajando en este ámbito. Asimismo, nueve de los diez sujetos habían cursado un máster relacionado con la traducción.

A los participantes se les envió un correo electrónico con toda la información relevante; en él, se les comunicaba en qué consistía cada prueba y cómo debían proceder y se les informaba de que los datos recogidos se procesarían de modo que fueran anónimos y solo se emplearían con una finalidad académica. Para la prueba de percepción, el MT Ranking, la instrucción más relevante era que debían escoger el segmento que considerasen que, a la hora de poseditar, requeriría un menor esfuerzo de posedición (es decir, aquel que requeriría el menor número de correcciones posible y que tardarían menos en poseditar). Para la prueba de productividad, se les explicaba que debían poseditar bajo el estándar de *equal to human quality* (es decir, *full post-*

editing). Tanto en la prueba de evaluación como en la prueba de productividad, los traductores recibieron un correo automático adicional desde la propia plataforma con un enlace a DQF que les permitía acceder.

Los sujetos dispusieron de dos semanas para realizar las dos pruebas; en primer lugar, los diez participantes efectuaron la prueba de MT Ranking, que permite evaluar varios motores de traducción automática sin que los evaluadores conozcan qué motor están analizando. De esta manera, se evita que los sujetos tengan prejuicios al decidir entre las distintas opciones que se les presentan. Al mismo tiempo, se realizaron las pruebas de evaluación de calidad en los niveles de fluidez y precisión. Posteriormente, un grupo de seis traductores, extraído de los diez primeros, efectuaron la prueba de productividad. Mediante esta, DQF recoge el tiempo empleado en cada segmento en milisegundos y el número de correcciones. Cada traductor de este segundo grupo realizó dos pruebas; así, los seis traductores poseditaron los dos tipos de texto (el de documentación y el de marketing) con un motor distinto. A los sujetos se les asignaron las pruebas y, en ellas, la tipología textual y el motor de traducción de forma aleatoria. Durante el tiempo en el que los participantes realizaron las pruebas no se intervino de ninguna manera en el experimento.

Análisis de los resultados

Para analizar los resultados obtenidos en las pruebas, se descargó toda la información proporcionada por la herramienta de DQF en archivos HTML y hojas CSV y se crearon dos hojas de datos con toda la información recopilada. En la primera, se volcó toda la información obtenida de las cuatro evaluaciones de calidad; se podía ver el segmento original, la traducción automática en bruto y la puntuación otorgada en fluidez y precisión. En la segunda, se añadieron todos los datos correspondientes a cada poseditor en cada una de las pruebas de productividad. Se prepararon dos hojas en un único libro: una para los resultados de las posediciones de marketing y otra para los de documentación del usuario.

Para la prueba de productividad también se creó una única hoja de datos en la que se aunaban todos los resultados obtenidos. Se empleó el programa R para realizar un análisis estadístico y extraer la media, la mediana, la desviación estándar y la distribución en cuartiles. A continuación, se efectuó un análisis comparativo entre pares de datos independientes. Para ello, primero se determinó si la variancia presente era probabilística a través de un test de Levene. Por último, se halló la interrelación entre los pares de datos con una t de Student.

Resultados

A continuación, se procede al análisis de los resultados obtenidos en las distintas pruebas. Para facilitar la comprensión, se analizan en primer lugar los resultados obtenidos en las pruebas de percepción y, posteriormente, los de las pruebas de productividad.

Resultados de las pruebas de percepción

En cuanto a la prueba de percepción, en 7 de cada 10 casos los participantes consideraron que tardarían menos en poseditar los resultados propuestos por el motor de TAN que aquellos que provenían del motor de TAE; en otras palabras, estimaron que, en la mayor parte de los casos, el motor de Google requeriría un número de correcciones menor a la hora de poseditar sus segmentos de traducción automática en bruto:

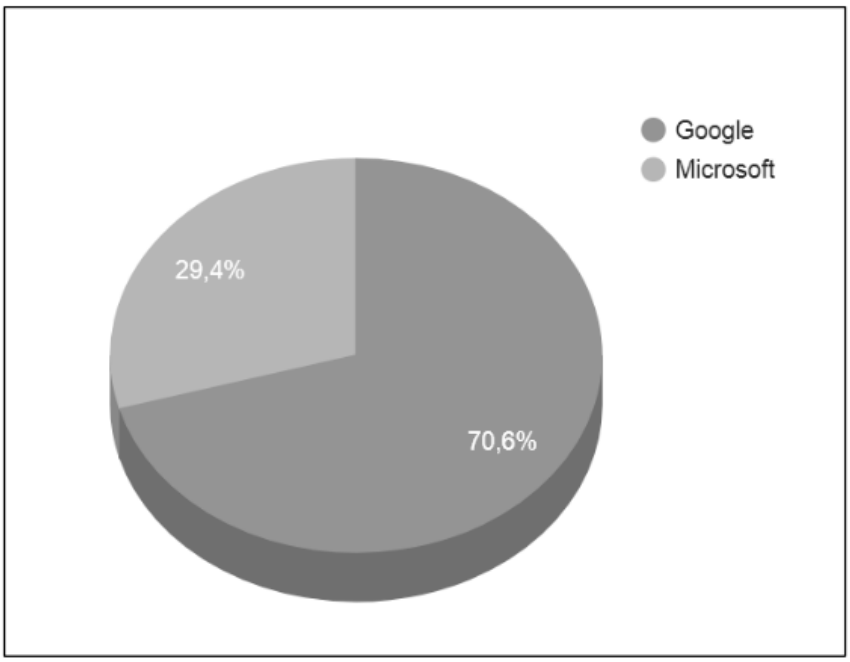

Figura 1: Datos de la prueba de MT Ranking

Por otra parte, aunque las pruebas de evaluación de calidad sirven de instrumento para analizar más profundamente los resultados de las pruebas de productividad, consideramos importante mencionar brevemente los resultados. En las siguientes gráficas se puede observar la comparación de la puntuación de los segmentos según las distintas categorías entre los segmentos propuestos por el motor de TAN y por el de TAE a nivel de fluidez y precisión:

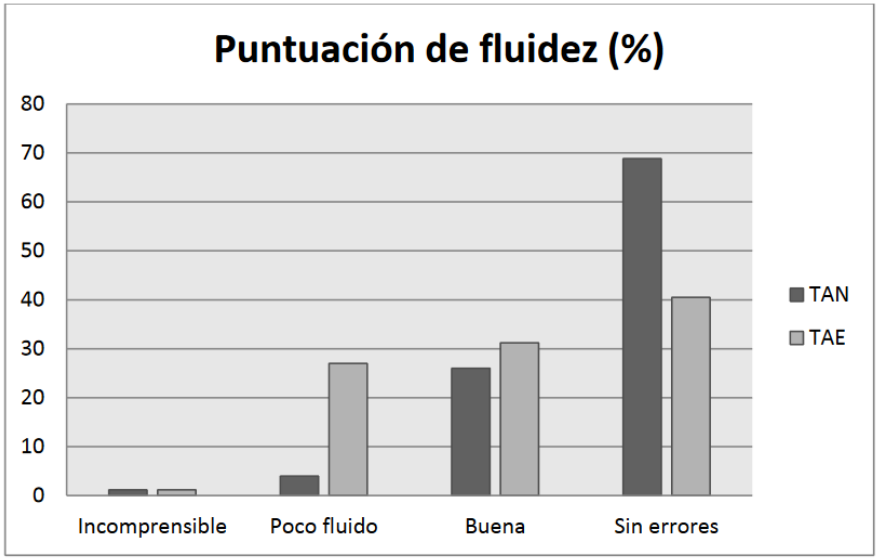

Figura 2: Datos de la prueba de fluidez

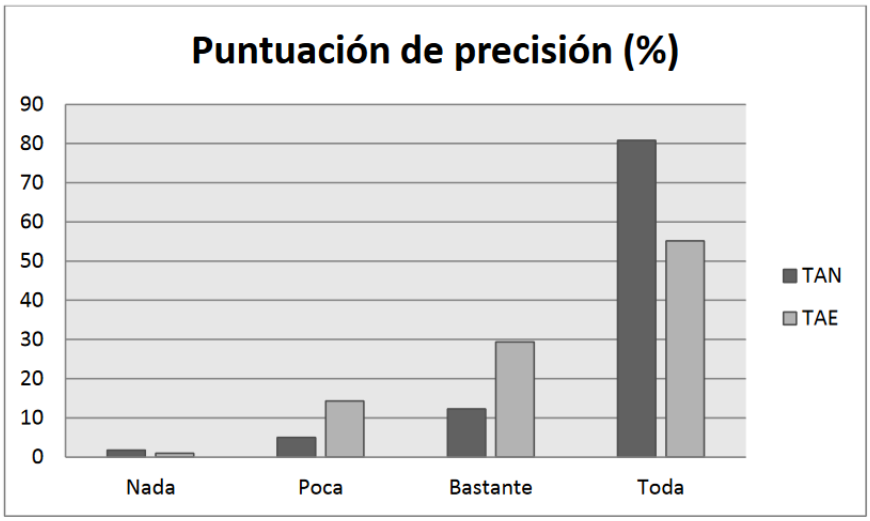

Figura 3: Datos de la prueba de precisión

Como se puede ver, si nos centramos en los niveles más altos de calidad tanto para fluidez como para precisión, el motor de TAN presenta unos resultados superiores al de TAE. No obstante, esto no sucede en la categoría intermedia. En ella, el motor de TAE muestra mejores resultados en el caso de la categoría «Bastante» en la evaluación de precisión. Sin embargo, en cuanto a la fluidez, se alcanzan resultados negativos, ya que la categoría «Poco fluido» presenta un porcentaje casi tres veces superior al porcentaje de los resultados del motor neuronal. Con respecto a las categorías de calidad más bajas, los dos motores indican unos resultados muy igualados.

Resultados de las pruebas de productividad

Los datos, en este caso, se categorizaron según su calidad (alta, media o baja) siguiendo los resultados obtenidos de la prueba de evaluación de calidad y según la longitud del segmento (menos de cinco palabras, de seis a diecinueve palabras y más de 20 palabras).

Aquí se presentan las medias obtenidas tanto de la distancia de edición como del tiempo de posedición en segundos. Los resultados se muestran resumidos en forma de tabla.

	Distancia de edición	Tiempo de posedición
Segmentos de menos de 5 palabras	TAN < TAE (p-valor < 0,05)	TAN > TAE (p-valor < 0,05)
Segmentos de 6 a 19 palabras	TAN < TAE (p-valor < 0,05)	TAN > TAE (p-valor > 0,05)
Segmentos de más de 20 palabras	TAN < TAE (p-valor > 0,05)	TAN > TAE (p-valor > 0,05)

Tabla 1: Resumen de los datos obtenidos para los segmentos de la categoría de calidad alta

En la categoría de mayor calidad, para los segmentos de menos de cinco palabras, la distancia de edición es significativamente mayor al poseditar los segmentos del motor de TAN (28,23) que al poseditar los de TAE (21,88), algo que sucede también con el tiempo de posedición (26,85 segundos para los resultados de TAN y 8,35 segundos para el de TAE, más de un tercio que el de TAN). Sin embargo, la muestra de datos obtenida en la distancia de edición es probabilística, algo que no sucede en la obtenida para el tiempo de posedición.

Para los segmentos de seis a diecinueve palabras de la categoría de mayor calidad, la distancia de edición de los resultados de media del motor de TAE (17,65) es significativamente mayor a la de TAN (13,63) con una muestra de datos probabilística. Aunque el tiempo de posedición es mayor al poseditar los segmentos de TAN (29,60 segundos) que al trabajar sobre los de TAE (19,63 segundos), no se puede afirmar que la diferencia de datos sea significativa.

En el caso de los segmentos de más de veinte palabras de mayor calidad, la distancia de edición es superior al poseditar los resultados obtenidos del motor de TAE (13,59) que al poseditar los del motor de TAN (11,59), pero esta vez no es una diferencia estadísticamente significativa. Esto mismo sucede con el tiempo de posedición, ya que, de media, se tarda 47,35 segundos en poseditar los resultados del motor neuronal y 32,54 segundos en poseditar los del motor estadístico. Así, sería necesario recabar más datos para evitar que la muestra sea tan heterogénea y conseguir datos estadísticamente significativos.

	Distancia de edición	Tiempo de posedición
Segmentos de menos de 5 palabras	TAN < TAE (p-valor < 0,05)	TAN > TAE (p-valor > 0,05)
Segmentos de 6 a 19 palabras	TAN < TAE (p-valor < 0,05)	TAN > TAE (p-valor > 0,05)

Segmentos de más de 20 palabras	TAN < TAE (p-valor > 0,05)	TAN > TAE (p-valor > 0,05)

Tabla 2: Resumen de los datos obtenidos para los segmentos de la categoría de calidad intermedia

Con respecto a la categoría de calidad intermedia, en el caso de los segmentos de menos de cinco palabras, la distancia de edición es significativamente menor cuando se usa el motor de TAN (27,16) que cuando se emplea el de TAE (52,91), de media, al contrario que en la categoría de calidad superior. Para este caso, el tiempo de posedición es considerablemente mayor al poseditar los resultados propuestos por el motor de TAN (45,19 segundos) que al poseditar los que propone el de TAE (19,65, menos de la mitad), pero, de nuevo, no de una forma estadísticamente significativa.

Para los resultados de los segmentos de entre seis y diecinueve palabras de esta categoría, al igual que en el caso anterior, la distancia de edición es significativamente menor al emplear el motor de TAN (29,96) que al utilizar el de TAE (15,35), con una t de Student con un valor de 1,079e-11. El tiempo de posedición vuelve a ser mayor al emplear el motor de TAN con respecto al de TAE, pero no de una forma estadísticamente significativa.

En el caso de los segmentos de más de veinte palabras, la distancia de edición vuelve a ser menor al poseditar los resultados del motor de TAN (19,75) que al trabajar sobre los del motor de TAE (26,33), pero no de una forma significativa. En cuanto al tiempo de posedición, es mayor al usar el motor de TAN (42,49 segundos) que al emplear el de TAE (36,51 segundos), pero no de una forma estadísticamente significativa.

En resumen, en la categoría intermedia, la distancia de edición es inferior en todos los casos, si bien esta no es significativa de forma estadística, de nuevo, en los segmentos de más de veinte palabras.

	Distancia de edición	Tiempo de posedición
Segmentos de menos de 5 palabras	TAN < TAE (p-valor < 0,05)	TAN > TAE (p-valor > 0,05)
Segmentos de 6 a 19 palabras	TAN < TAE (p-valor < 0,05)	TAN < TAE (p-valor > 0,05)
Segmentos de más de 20 palabras	TAN < TAE (p-valor < 0,05)	TAN > TAE (p-valor > 0,05)

Tabla 3: Resumen de los datos obtenidos para los segmentos de la categoría de calidad baja

En la categoría de calidad baja, para los segmentos de menos de cinco palabras la distancia de edición de los resultados del motor de TAN es, una vez más, inferior de forma estadísticamente significativa. En cuanto al tiempo de posedición, es mayor para los segmentos poseditados con el motor de TAN que para los poseditados con el motor de TAE, pero no de una forma estadísticamente significativa.

En los segmentos de entre seis y diecinueve palabras de esta categoría, al igual que en los de menos de cinco palabras, la distancia de edición es significativamente inferior al poseditar los resultados ofrecidos por el motor de TAN que al trabajar sobre aquellos obtenidos con el de TAE, con una t de Student que arrojó un valor muy bajo (< 2,2e-16). No obstante, el tiempo de posedición es inferior al emplear el motor de TAN que al usar el de TAE, a diferencia de lo que sucede en el caso de los segmentos de menos de cinco palabras (de nuevo, sin embargo, no de una forma estadísticamente significativa).

En cuanto a los segmentos de más de veinte palabras, la distancia de edición es inferior, una vez más, al emplear el motor de TAN que al utilizar el de TAE de forma estadísticamente significativa. El tiempo de posedición, por su parte, vuelve a ser mayor al emplear el motor de TAN (54,92 segundos) que al usar el de TAE (51,60 segundos), pero los valores no son estadísticamente significativos.

En total, de los nueve casos posibles presentados, en ocho de ellos la distancia de edición es inferior al poseditar los resultados obtenidos por parte del motor de TAN que al poseditar los del motor de TAE. Asimismo, únicamente en dos de estos escenarios los datos no son estadísticamente significativos: cuando los segmentos tienen más de veinte palabras en la categoría de calidad alta y cuando los segmentos tienen de seis a diecinueve palabras en la categoría de calidad intermedia. Sobre el tiempo de posedición, en todos los casos los sujetos necesitaron más tiempo para realizar la posedición de los segmentos de traducción neuronal que para poseditar los de traducción estadística. Sin embargo, la muestra obtenida en estos casos es demasiado heterogénea y los resultados no son estadísticamente significativos.

Conclusiones

Para determinar la percepción de los traductores, se llevó a cabo una prueba de evaluación de motores (MT Ranking) en la plataforma DQF de TAUS. En ella, diez traductores escogieron entre dos segmentos de destino de traducción automática en bruto, uno de un motor neuronal (Google Neural Machine Translation) y otro de un motor estadístico (Microsoft Translator), para un mismo segmento de origen. Debían escoger el que ellos considerasen que sería más productivo en caso de poseditarlo; esto es: el que requeriría menos correcciones y tardarían menos en poseditar. Así, en el 70 % de los casos, los participantes consideraron que el motor de TAN sería más productivo que el de TAE. Por tanto, se confirma la hipótesis planteada al inicio de que la posedición de los segmentos de TAN es percibida como un proceso más productivo. Esto puede deberse a la suposición por parte de los traductores de que los grandes avances propiciados por la irrupción de la traducción automática neuronal en el mercado garantizan también una mejor calidad en los segmentos propuestos y un menor tiempo de posedición al trabajar con ellos.

Sin embargo, los resultados obtenidos en las pruebas de percepción contrastan con los recabados en las pruebas de productividad: aunque el motor neuronal es considerado más productivo, el análisis de los resultados de las pruebas de productividad llevadas

a cabo por seis de estos diez traductores no corrobora esta apreciación. Si bien la distancia de edición es menor en los segmentos poseditados con el motor de TAN que con el de TAE, el esfuerzo temporal es mayor al emplear el motor neuronal. Por tanto, la percepción de los traductores no se corresponde en su totalidad con la realidad, ya que sí es cierto que son necesarias menos modificaciones, pero el tiempo para llevarlas a cabo es mayor. Se confirma así la hipótesis que se presentaba al inicio de que existen diferencias en cuanto a esfuerzo temporal y técnico en los resultados de la posedición de los segmentos de los motores de TAN y TAE. Asimismo, se extrae la conclusión de que, aunque los traductores consideran que el motor neuronal es de mejor calidad porque puede que haya menos errores en los segmentos que propone, estos son mucho más difíciles de detectar, ya que se emplea más tiempo en poseditar estos segmentos que en poseditar los del motor estadístico.

En cuanto a las conclusiones obtenidas de los resultados de las pruebas de evaluación de calidad, los datos muestran que el motor neuronal presenta una fluidez y una precisión mayores que las del motor de TAE. De aquí se puede deducir que, de la misma forma que en las pruebas de percepción, el evaluador considera más adecuados, en términos de fluidez y precisión, los resultados del motor de traducción automática neuronal. Por tanto, el sujeto presupone que estos resultados requerirán menos correcciones y menos tiempo de posedición. Sin embargo, en este caso ya ha quedado demostrado que le llevaría más tiempo poseditarlos.

Futuras líneas de investigación

En cuanto al trabajo futuro en esta área, en primer lugar, consideramos que sería necesario repetir las pruebas de productividad de cara a obtener una muestra más homogénea de datos relacionados con el esfuerzo temporal. De esta forma, se podría repetir el análisis estadístico y así obtener unos datos estadísticamente significativos.

Otra de las líneas de investigación que sería interesante seguir consistiría en replicar las pruebas con las mismas propuestas de segmentos de traducción automática estadística de Microsoft y con el propio motor de traducción automática neuronal de esta misma empresa. Siguiendo con la replicabilidad, sería importante comprobar si se obtendrían los mismos resultados con tipologías textuales diferentes: por ejemplo, con textos legales. Asimismo, sería interesante observar qué resultados se obtienen al realizar de nuevo las pruebas con otras combinaciones lingüísticas. Esto podría hacerse con idiomas próximos entre ellos (como el español y el catalán) o con lenguas con las que se conoce que los motores de traducción funcionan bien (como en el par de idiomas inglés y alemán).

Por otra parte, valdría la pena continuar trabajando con los resultados obtenidos a partir de las pruebas de calidad, de un modo similar al estudio llevado a cabo por Esperança-Rodier *et al.* (2017). Se podría establecer una categorización de errores dentro del marco de MQM para detectar qué número de errores y qué gravedad le corresponden a cada propuesta del motor y determinar si, de esta forma, los datos de productividad obtenidos en las pruebas de productividad se ven justificados.

Bibliografía

Aranberri, N. (2014). Posedición, productividad y calidad. *Tradumàtica: Tecnologies de la Traducció,* n. 12, pp. 471–477. <http://revistes.uab.cat/tradumatica/article/view/n12-aranberri>. <https://doi.org/10.5565/rev/tradumatica.62>

Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas: Association for Computational Linguistics, pp. 257-267. <https://www.aclweb.org/anthology/D16-1025.pdf. https://doi.org/10.18653/v1/D16-1025>.

Esperança-Rodier, E.; Rossi, C.; Bérard, A.; Besacier, L. (2017). Evaluation of NMT and SMT Systems: A Study on Uses and Perceptions, in: *Proceedings of the 39th Conference Translating and the Computer*. London: AsLing, pp. 11-24. <https://www.semanticscholar.org/paper/Evaluation-of-NMT-and-SMT-Systems%3A-A-Study-on-Uses-Esperan%C3%A7a-Rodier-Berard/58b0ffc3892f0f24594aa424a60de0f18aab970a>.

Görög, A. (2014). Quality evaluation today: the Dynamic Quality Framework, in: *Proceedings of Translating and the Computer* 36: ASLING: Proceedings. Geneva: Tradulex, pp. 155-164. <http://www.tradulex.com/varia/TC36-london2014.pdf>.

Guerberof, A. (2009). Productivity and quality in MT post-editing, in: *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII), Beyond Translation Memories: New Tools for Translators Workshop.* Ottawa, Canadá: MT Summit. <http://www.mt-archive.info/MTS-2009-Guerberof.pdf>.

Guerberof, A. (2013). What do professional translators think about post-editing? *The Journal of Specialised Translation,* n. 19, pp. 75–95. <http://www.jostrans.org/issue19/art_guerberof.php>.

Katan, D. (2016). Translation at the cross-roads: Time for the transcreational turn? *Perspectives. Studies in Translatology*, v. 24, n. 3, pp. 365–381. <https://doi.org/10.1080/0907676X.2015.1016049>

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations, in: *Proceedings of the 7th Workshop on Statistical Machine Translation*. Montreal, Canadá: Association for Computational Linguistics, pp. 181-190. <https://www.aclweb.org/anthology/W12-3123>.

Krings, H. P. (2001). *Repairing texts*. Kent: Kent State University Press.

Lohar, P.; Popovic, M.; Afli, H.; Way, A. (2019). A Systematic Comparison Between SMT and NMT on Translating User-Generated Content, in: *Proceedings of CICLing 2019, the 20th International Conference on Computational Linguistics and Intelligent Text Processing, La Rochelle, France.* <https://www.computing.dcu.ie/~away/PUBS/2019/A_Systematic_Comparison_Between_SMT_and_NMT_on_Translating_User_Generated_Content.pdf>

Moorkens, J. (2017). Under pressure: translation in times of austerity. *Perspectives,* v. 25, n. 3, pp. 464-477. <https://doi.org/10.1080/0907676X.2017.1285331>.

Moorkens, J.; Toral, A.; Castilho, S.; Way, A. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces,* v. 7, n. 2, pp. 240-262. <https://www.rug.nl/research/portal/publications/translators-perceptions-of-literary-postediting-using-statistical-and-neural-machine-translation(af141520-ca42-4b37-a779-a5f4b829f39e).html. https://doi.org/10.1075/ts.18014.moo>

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation, in: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Filadelfia: Association for Computational Linguistics, pp. 311-318. <https://doi.org/10.3115/1073083.1073135>.

Shterionov, D.; Nagle, P.; Casanellas, L.; Superbo, R.; O'Dowd, T. (2017). Empirical evaluation of NMT and PBSMT quality for large-scale translation production, in: *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT): User Track*. Praga: European Association of Machine Translation, pp. 74-79. <https://ufal.mff.cuni.cz/eamt2017/user-project-product-papers/papers/user/EAMT2017_paper_83.pdf>.

Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. (2006). A study of translation edit rate with targeted human annotation, in: *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006).* Cambridge, Massachusetts: Association for Machine Translation in the Americas.

Torres-Hostench, O.; Presas, M.; Cid-Leal, P. (2016). *El uso de traducción automática y posedición en las empresas de servicios lingüísticos españolas: informe de investigación ProjecTA 2015.* Bellaterra, Cerdanyola del Vallès. <https://ddd.uab.cat/record/166753>.

Vilar, D.; Xu, J.; D'Haro L. F.; *et al.* (2006). Error analysis of statistical machine translation output, in: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association*. Genoa, Italia, pp. 697-702. <http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf>.

Way, A. (2013). Traditional and Emerging Use-Cases for Machine Translation, in: Proceedings of Translating and the Computer 35. London. <https://www.computing.dcu.ie/~away/PUBS/2013/Way_ASLIB_2013.pdf>.

Wołk, K.; Koržinek, D. (2017). Comparison and Adaptation of Automatic Evaluation Metrics for Quality Assessment of Re-Speaking. *Computer Science*, v. 18, n. 2, pp. 129. <https://doi.org/10.7494/csci.2017.18.2.129>.