Traducción automática inglés-catalán:

tecnología de vanguardia, calidad y productividad

Trabajo de Fin de Máster

Vicent Briva-Iglesias
Director: Adrià Martín-Mor

Universitat Autònoma de Barcelona

Resumen
Los grandes cambios y avances tecnológicos recientes han consolidado la traducción automática (TA) como un elemento clave a tener en cuenta en el mundo de los servicios lingüísticos. En muchos casos, es incluso un actor esencial por las limitaciones de presupuesto y tiempo. En los últimos años se ha prestado mucha atención a la investigación sobre la TA y el uso de usuarios profesionales o aficionados ha aumentado. Sin embargo, la investigación se ha centrado principalmente en las combinaciones de idiomas con grandes cantidades de corpus disponibles en línea (por ejemplo, inglés-español). La situación de las lenguas minoritarias o sin estado como el catalán es diferente. En este estudio se analiza el nuevo motor de traducción automática neuronal de código libre de Softcatalà y se compara con Apertium y el Traductor de Google en la combinación inglés-catalán. Aunque los desarrolladores de motores de TA utilizan métricas automáticas para la evaluación de los motores, la evaluación humana sigue siendo la referencia a pesar de su coste elevado. Con DQF de TAUS, se ha evaluado la calidad de traducción (en términos de clasificación relativa, precisión y fluidez) y la productividad (al comparar los tiempos y las distancias de edición) con la participación de 11 personas evaluadoras. Los resultados muestran que el Traductor de Softcatalà ofrece una calidad y una productividad superior a las del resto de motores analizados.

Palabras clave
Traducción automática; traducción automática neuronal; evaluación humana; calidad de traducción; código libre.

Resum
Els grans canvis i avenços tecnològics recents han consolidat la traducció automàtica (TA) com un element clau a tenir en compte en el món dels serveis lingüístics. En molts casos, és fins i tot un actor essencial per les limitacions de pressupost i temps. En els últims anys s'ha prestat molta atenció a la investigació sobre la TA i l'ús d'usuaris professionals o aficionats ha augmentat. No obstant això, la investigació s'ha centrat principalment en les combinacions d'idiomes amb grans quantitats de corpus disponibles en línia (per exemple, anglès-espanyol). La situació de les llengües minoritàries o sense estat com el català és diferent. En aquest estudi s'analitza el nou motor de traducció automàtica neuronal de codi lliure de Softcatalà i es compara amb Apertium i el Traductor de Google a la combinació anglès-català. Tot i que els desenvolupadors de motors de TA utilitzen mètriques automàtiques per a l'avaluació dels motors, l'avaluació humana segueix sent la referència malgrat el seu cost elevat. Amb DQF de TAUS, s'ha avaluat la qualitat de traducció (en termes de classificació relativa, precisió i fluïdesa) i la productivitat (en comparar els temps i les distàncies d'edició) amb la participació d'11 persones avaluadores. Els resultats mostren que el Traductor de Softcatalà ofereix una qualitat i una productivitat superior a les de la resta de motors analitzats.

Paraules clau
Traducció automàtica; traducció automàtica neuronal; avaluació humana; qualitat de traducció; codi lliure.

Abstract
Recent major changes and technological advances have consolidated machine translation (MT) as a key player to be taken into account in the language services world. In many cases, it is even an essential player due to budget and time constraints. Much attention has been paid to MT research in recent years, and MT use by professional or amateur users has increased. Yet, research has focused mainly on language combinations with huge amounts of online available corpora (e.g. English-Spanish). Nevertheless, the situation for minoritized or stateless languages like Catalan is different. This study analyses Softcatalà's new open-source, neural machine translation engine and compares it with Apertium and Google Translator in the English-Catalan combination. Although MT engine developers use automatic metrics for MT engine evaluation, human evaluation remains the gold standard despite its cost. Using TAUS DQF tools, translation quality (in terms of relative ranking, accuracy and fluency) and productivity (comparing editing times and distances) have been evaluated with the participation of 11 evaluators. Results show that Softcatalà's Translator offers higher quality and productivity than the other engines analysed.

Keywords
Machine translation; neural machine translation; human evaluation; translation quality; open source.

Índice

Índice de figuras

Índice de tablas

1. Introducción

Según Schwab (2016, pp. 11-16), actualmente nos encontramos en la Cuarta Revolución Industrial, basada en las tecnologías digitales. Al igual que en las revoluciones industriales previas, los cambios abruptos y el fenómeno de la globalización han creado nuevas oportunidades, pero también daños y polarización en las economías y las sociedades modernas (Schwab, 2019, p. 7). El ejercicio de la traducción no es una excepción. Los grandes cambios y avances tecnológicos recientes han consolidado la TA como un actor clave y a tener en cuenta en el proceso de los servicios lingüísticos (traducción, localización, internacionalización, etc.). En muchos casos, incluso en un actor imprescindible. Esto ha provocado cambios continuos, generalizados y profundos en el mundo profesional y académico de la traducción (Cronin, 2012).

Desde la creación de los primeros ordenadores, surgió la idea de automatizar el lenguaje y se empezó a teorizar sobre traducción automática (Shannon y Weaver, 1949), ya que mucha gente creyó que las barreras lingüísticas caerían en poco tiempo gracias a esta tecnología (Hutchins, 1986). Esto provocó que grandes empresas tecnológicas y gobiernos —como, por ejemplo, el de EE. UU., con la creación del *Automatic Language Processing Advisory Comittee* (ALPAC)— invirtieran en ella. No obstante, poco después, en 1966, el informe de este mismo comité mostraba claramente que el lenguaje y la comunicación son tal vez una de las barreras más difíciles de saltar para la tecnología y que aún queda un gran camino, en caso de que esto suceda: «there is no immediate or predictable prospect of useful machine translation» (ALPAC, 1966, p. 32), ya que la TA era un proceso más caro que la traducción humana (TH) y no cumplía con las expectativas de calidad.

Con los avances tecnológicos actuales, el uso de la inteligencia artificial y el desarrollo del aprendizaje profundo o *Deep Learning*, la potencia de los ordenadores y de las máquinas ha aumentado exponencialmente, lo que nos ha permitido reducir el tiempo de procesamiento de muchas tareas automáticas y aumentar su productividad (Ahrenberg, 2017). En este contexto, ha aparecido la traducción automática neuronal (TAN), un nuevo sistema de traducción automática mucho más

complejo que se cree que acabará con los sistemas o paradigmas anteriores (Bentivogli et al., 2016). Según Wu et al. (2016, p. 2), «the quality of the resulting [neural machine] translation system gets closer to that of average human translators». Muchos otros investigadores también afirman que la TAN ha superado al paradigma de TA anterior, la traducción automática estadística (TAE) (Bahdanau et al., 2016; Ondřej Bojar et al., 2016; Jean et al., 2015). No obstante, estos estudios enuncian una calidad de traducción mayor mediante un análisis con métricas automáticas o ligeramente humanas. Según Castilho et al. (2017), la TAN es un avance en el campo de la TA, pero debe tenerse cuidado, ya que aún presenta ciertas limitaciones palpables al hacer una evaluación humana.

2. Justificación y objetivos

Mi gran interés por la tecnología y su aplicación al flujo de trabajo profesional o académico hacen que quiera indagar más en la utilización y la evaluación de motores de TA en sus versiones y modos de vanguardia, como por ejemplo la nueva TAN. A esto se suma mi interés por las herramientas de código libre y las combinaciones de idiomas menos investigadas, como puede ser el inglés-catalán. Dicho esto, el objetivo del presente trabajo es ahondar en el conocimiento de los motores de TA y su evaluación en la combinación inglés-catalán. Asimismo, este estudio también pretende analizar la calidad y el rendimiento de un nuevo motor de TAN de código libre, creado por Softcatalà, y compararlo con su predecesor de código libre, Apertium —un motor de traducción automática basada en reglas (TABR)—, y el motor de TAN privativo de Google. Se han escogido estos motores por diversos motivos. En primer lugar, el Traductor de Google es el motor comercial y privativo más conocido y utilizado y, por tanto, es el que utiliza la mayoría de los estudiantes o de las personas no profesionales que utilizan sistemas de TA. En segundo lugar, Apertium es el motor que más utilizan los miembros de Softcatalà en sus colaboraciones de localización de software libre y han demostrado su interés en saber si vale la pena cambiar de sistema de TA para aumentar su productividad. Por lo tanto, Mi investigación pretende conseguir lo siguiente:

1. Revisar la bibliografía más actualizada sobre motores de TA y su evaluación.
2. Analizar qué método y sistema de TA escogido, ya sea Apertium (TABR de código libre), el Traductor de Softcatalà (TAN de código libre) o el Traductor de Google (TAN de código privativo), ofrece una calidad de traducción mayor.
3. Descubrir qué motor de TA escogido ofrece un rendimiento mayor al poseditar, con el objetivo de saber cuál puede aportar más beneficios al introducirlo en un flujo de traducción profesional o aficionado.
4. Sentar las bases para una investigación futura y un posible doctorado, ya sea en la creación de motores de TA en dominios específicos (jurídica, médica, etc.) o en la formación de traductores (la utilidad de la TA en la formación; cómo introducir la TA en el flujo profesional o académico, en los servicios públicos, etc.).

3. Contextualización del objeto de estudio

A la hora de definir el marco teórico, es imprescindible presentar el contexto en el que se desarrolla el presente estudio. Para ello, hay que hacer un breve repaso de las tecnologías de la traducción. A continuación, se hará una pequeña descripción de la TA, su definición, los tipos de motores disponibles en la actualidad y para qué se utiliza. Asimismo, se delimitará qué es la calidad de traducción y su evaluación, para así poder comparar los motores escogidos. Finalmente, se tratará la posedición, ya que es el fin más común y práctico de la TA.

3.1. Las «nuevas» tecnologías de la traducción

Pese a que el término tecnología parece relacionarse con los dispositivos móviles y la nube actuales, las tecnologías aplicadas al mundo profesional de la traducción no son tan «nuevas». Para hablar de ellas, cabe remontarse de nuevo al informe del *Automatic Language Processing Advisory Comittee* (ALPAC, 1966), mencionado anteriormente en el apartado 2. Justificación y objetivos. El fracaso de la incipiente TA en la segunda mitad del siglo XX y la recomendación que el informe hacía sobre que el interés en la traducción automática tenía que transferirse a la traducción asistida por las máquinas, la cual estaba orientada a «improve human translation, with an appropriate use of

machine aids» (ibid.:25), fue la semilla que germinó en lo que hoy conocemos como tecnologías de traducción.

En primer lugar, surgió por primera vez la idea de recuperar automáticamente segmentos de texto que ya se habían traducido previamente con el objetivo de facilitar el proceso de traducción (Arthern, 1979; Melby, 1978). Posteriormente, hubo varias empresas que empezaron a interesarse por el desarrollo de software relacionado con los procesos traductológicos, entre las que destacan Trados y Star Group, pioneras en el campo. A causa de los avances tecnológicos, los procesos de traducción y documentación profesionales se han visto alterados por la introducción de herramientas creadas con la intención de traducir el mismo contenido en menos tiempo y aumentar la productividad. Es aquí cuando empiezan a aparecer las fuentes de información en línea —diccionarios, enciclopedias, corpus lingüísticos en varios idiomas, etc.— que se utilizan para sustituir los recursos en papel que se utilizaban previamente. Además, también aparecen las primeras herramientas de gestión de terminología como Multiterm, con la intención de poder mostrar las equivalencias terminológicas de varios idiomas de manera automática y más rápida (Chan, 2014). Todas estas herramientas han acabado sumándose y formando lo que lo que hoy llamamos herramientas de traducción asistida por ordenador (TAO) o, de manera más concisa, sistemas de gestión y edición de traducciones (SGET) (Martín-Mor et al., 2016). Estos sistemas almacenan los segmentos ya traducidos y funcionan como memorias de traducción, entre otras funciones, y permiten trabajar con una gran variedad de archivos. En cuanto a los SGET, hay de software privativo como SDL Trados Studio, memoQ y Memsource, o de software libre como OmegaT o Matecat.

3.2. La traducción automática (TA)

Pese a que la investigación en traducción automática se abandonó ligeramente y se apartó de las prioridades tras el informe del ALPAC, continuó habiendo investigación al respecto, aunque en menor medida. Después de la Segunda Guerra Mundial, la invención de los ordenadores y los avances en teoría de la información permitieron pensar de nuevo que la traducción automática era un hecho plausible. Pero para hablar

de traducción automática, primero hay que saber qué se entiende cuando se utiliza el término. Según (Ginestí y Forcada, 2009, p. 43):

> La traducció automàtica (TA) és el procés de traducció, mitjançant un sistema informàtic (compost per ordinadors i programes), de textos informatitzats escrits en la llengua origen a textos informatitzats escrits en la llengua meta. Un text informatitzat és un fitxer d'ordinador que conté un text en algun format conegut.

Por lo tanto, podemos decir que la TA es una herramienta que, cuando se introduce un texto informatizado en un lenguaje natural, traduce de manera automática el texto original a otro lenguaje natural, produciendo lo que se llama traducción en bruto. Históricamente, la TA se ha basado en dos grandes tipos, que explicaremos a continuación: la TA basada en reglas y la TA basada en corpus (Nitzke, 2019).

3.2.1. La traducción automática basada en reglas (TABR)

La TABR fue el primer sistema de TA en desarrollarse, el cual se centra en traducir palabra por palabra (ibid.: 6-8). Este primer sistema de TA requiere de dos actores principales para que la traducción fuese posible:

1. En primer lugar, un conjunto de lingüistas o expertos en traducción, que tienen que definir una serie de diccionarios y normas lingüísticas (gramaticales, sintácticas y estilísticas) de las características de la lengua origen (LO) y de la lengua meta (LM).
2. En segundo lugar, también se necesitan expertos informáticos que serán los que escribirán programas analizadores de morfología y sintaxis capaces de leer, entender y representar las reglas del diccionario con el objetivo de aplicarlas en la LM. Estos expertos informáticos también pueden ser lingüistas computacionales o lingüistas con conocimientos de código.

Este sistema puede ser muy útil para los idiomas minorizados y con pocos textos en línea disponibles y ha demostrado ofrecer una buena calidad de traducción en varias combinaciones de idiomas, especialmente en aquellas de idiomas cercanos (como, por

ejemplo, el español y el catalán). No obstante, los lenguajes son complejos y la ambigüedad léxica o sintáctica es uno de los problemas más difíciles de solventar (Ginestí y Forcada, 2009). Estas ambigüedades pueden afrontarse mediante la creación de reglas nuevas. Por ejemplo, el término «sheet» puede traducirse como «sábana» o «hoja». Se pueden definir reglas de contexto como la siguiente:

```
<rule>
  <or>
   <match lemma="bed" tags="n.*"/>
   <match lemma="woolen" tags="adj"/>
  </or>
  <match lemma="sheet" tags="n.*">
   <select lemma="sábana" tags="n.*"/>
  </match>
</rule>
```

En este contexto, se ha añadido una condición. Si el término «sheet» aparece acompañado de «bed» o «woolen», se traducirá por «sábana» en lugar de por «hoja». El problema de este sistema es que el coste de creación y compilación de las reglas es muy elevado, ya que para conseguir un buen sistema de TABR hay que escribir muchas reglas y esto implica un trabajo arduo y complejo, ya que los idiomas tienen muchas excepciones e implicaciones que son difíciles de representar con reglas por escrito (Moorkens et al., 2018). Algunos motores de TABR son Apertium (de código libre), que estudiaremos y analizaremos en el presente trabajo, o Lucy (de código privativo).

3.2.2. La traducción automática estadística (TAE)

Con la proliferación de Internet y la globalización, hay una gran cantidad de textos disponibles en línea en muchos idiomas diferentes que pueden utilizarse libremente. La TAE es un tipo de traducción automática basada en corpus. Esto significa que es un sistema informático que aprende a traducir automáticamente a partir de traducciones existentes alineadas en forma de corpus (Hearne y Way, 2011). Los sistemas de TAE pasan por dos procesos diferentes.

1. El **entrenamiento**: este primer proceso implica la extracción de un modelo de traducción estadístico a partir de un corpus paralelo (y así asociar las probabilidades de que una palabra en la LO se traduzca de determinada manera en la LM). Además, el entrenamiento también implica la creación de un modelo de lengua a partir de un corpus monolingüe (es decir, un diccionario probabilístico que permite estimar la fluidez de un texto en la LM) (Brown et al., 1990).
2. El **«decoding»**: este segundo proceso afronta la traducción como un problema de búsqueda. Cuando se introduce una oración en la LO en el motor de TAE, este busca todas las traducciones posibles según el modelo de traducción. A continuación, intenta reordenar estas traducciones de conformidad con el modelo de lengua y, para acabar, ofrece como traducción en bruto la oración en la LM que sea «más probable».

Según (Hearne y Way, 2011):

> SMT is, however, probabilistically plausible; rather than focusing on the best process to use to generate a single optimal translation for a source sentence, SMT focuses on generating many thousands of hypothetical translations for the input string, and then working out which one of those is most likely.

Con el objetivo de ilustrar este proceso de una manera sencilla, esta sería una simplificación del proceso:

En primer lugar, los textos deben segmentarse en oraciones. Los signos de puntuación se utilizan como marcadores que indican el final de una oración, así se reconoce de manera automática qué se debe separar y qué es una oración independiente. Por ejemplo, si tenemos el texto «El perro es verde. El gato es marrón», el punto se utilizará como elemento separador de dos oraciones y obtendremos, por una parte, «El perro es verde», y por otra, «El gato es marrón». Lo mismo pasará con la versión en inglés «The dog is green. The cat is brown».

En segundo lugar, los textos deben alinearse. Para que el sistema funcione correctamente, es necesario que las oraciones en la LO estén alineadas con sus homónimas en la LM. Por lo tanto, deberán alinearse «El perro es verde» con «The dog is green». Así como «El gato es marrón» con «The cat is brown». Esta alineación de todas las oraciones de unos textos bilingües —o corpus— se hace con un único objetivo: crear el modelo de traducción y así saber qué palabra en la LM corresponde a qué palabra en la LO (Brown et al., 1990). Una vez alineadas las oraciones, tienen lugar una serie de operaciones probabilísticas explicadas a continuación, con la intención de descubrir las probables traducciones de los términos de las oraciones:

The dog is green	The cat is brown	The dog
El perro es verde	El gato es marrón	El perro

Primero, tiene lugar un proceso llamado inicialización. En este punto, todas las palabras tienen la misma probabilidad de alinearse unas con otras. Es decir, en la primera oración, «The dog is green», la palabra «The» podría corresponderse con «El», «perro», «es» o «verde». No obstante, gracias a un proceso de iteración, en el que se comparan las probabilidades con las oraciones contiguas, al ver que «The dog is green» se corresponde a «El perro es verde», que «The cat is brown» se corresponde a «El gato es marrón» y que «The dog» se corresponde a «El perro», podemos ir viendo cómo la probabilidad de que una palabra sea la traducción de otra aumenta. Tras hacer esta comparación y este análisis probabilístico sencillo, observamos que «The» es el equivalente en LO de «El» en LM, así como que «dog» va de la mano de «perro», y «cat» de «gato», por decir algunos ejemplos. Este último proceso de alineación palabra a palabra se llama convergencia. Gracias a estos alineamientos, podemos obtener un diccionario bilingüe probabilístico, es decir, obtenemos las probabilidades de que «cat» sea «gato» —expresado en [p(cat|gato)]—, consiguiendo así nuestro objetivo: un modelo de traducción. A estas *p* (probabilidad) se le asigna un peso, que aumentará o disminuirá según la probabilidad de que la palabra en LO sea correcta en LM. No obstante, se debe tener en cuenta un elemento adicional: la corrección o la naturalidad de las oraciones en la LM.

A este fin, el motor de TAE debe entrenarse con un texto monolingüe en la LM, es decir, en vez de hacer un análisis probabilístico de un corpus bilingüe para obtener las probabilidades de traducción (como ya se ha hecho para obtener el modelo de traducción), debe hacerse un análisis probabilístico del texto en la LM para predecir el orden natural y marcar la norma de las palabras y de los segmentos. Tras este proceso, obtendremos el «modelo de lengua», que marcará y fijará la norma lingüística. Como con los corpus bilingües, cuanto más grande sea y cuantas más oraciones y segmentos tenga, mejor será el resultado. Asimismo, a este modelo de traducción y al modelo de lengua se les pueden sumar otros modelos diferentes, como el de reordenamiento o el de traducción de segmentos, que condicionarán las probabilidades y los pesos de las palabras según cuál se utilice. No obstante, no entraremos a valorarlos en detalle, por las limitaciones del trabajo.

En este momento, el proceso que sigue es el «decoding». En esta etapa, el sistema de TAE busca todas las posibles hipótesis de traducción y las valora según su puntuación y peso final. Aquella hipótesis con una puntuación mayor (teniendo en cuenta tanto el modelo de lengua como el de traducción de manera conjunta), será la opción elegida por el motor de TAE y la que se mostrará como texto en bruto. Algunos ejemplos de motores de TAE son Moses (Koehn et al., 2007) y MTradumática (Martín-Mor, 2017).

3.2.3. La traducción automática neuronal (TAN)

Recientemente, un gran desarrollo de la TAE y la adopción de redes neuronales ha permitido crear la TAN. Según (Koehn, 2017), los modelos de redes neuronales ya se habían propuesto de manera muy similar anteriormente (Castaño et al., 1997; Forcada y Ñeco, 1997). No obstante, ninguno de estos modelos se había podido entrenar con cantidades de texto como las que están disponibles ahora mismo y, además, la potencia de procesamiento de las máquinas y los ordenadores de entonces no era la de ahora. Por tanto, solo los avances en procesamiento actuales han permitido «resucitar» los antiguos modelos neuronales, que se han abierto hueco en el mundo de la TA. Forcada (2017, p. 292) define la TAN de la siguiente manera:

> The name comes from the fact that the neural networks (which should properly be called artificial neural networks) on which NMT is based are composed of thousands of artificial units that resemble neurons in that their output or activation (that is, the degree to which they are excited or inhibited) depends on the stimuli they receive from other neurons and the strength of the connections along which these stimuli are passed.

Además, también indica que los recursos materiales necesarios para crear motores y poder utilizar este nuevo sistema de TA no son tan accesibles ni tan fáciles de conseguir como los requeridos en la TAE. La TAN requiere de corpus lingüísticos mucho más grandes que la TAE, para poder así «estimular» las neuronas adecuadamente durante la fase del entrenamiento del motor. Además, no pueden hacerse motores con ordenadores convencionales, sino que se requieren procesadores muy potentes llamados unidades de procesamiento gráfico o GPU. Al ser un proceso muy exigente, el entrenamiento puede tardar días o incluso meses. Hasta hace poco, la TAE era el sistema más utilizado en el mundo de la traducción profesional. No obstante, algunos estudios ya han demostrado que la TAN parece tener un rendimiento superior (Bojar et al., 2016) y la industria se ha volcado a introducir y actualizar sus sistemas de TA, como por ejemplo Google (Wu et al., 2016), Systran (Crego y Senellart, 2016) y la OMPI (Junczys-Dowmunt y Grundkiewicz, 2016). Sin embargo, y tal y como se había comentado en el apartado 2. Justificación y objetivos, la TAN es un sistema con muchas posibilidades de superar al paradigma anterior, pero aún es relativamente nuevo y debe analizarse y estudiarse con cuidado antes de hacer presunciones o afirmaciones (Castilho et al., 2017). Algunos de los motores de TAN más destacados son DeepL y algunas combinaciones de idiomas del Traductor de Google.

3.2.4. La traducción automática interactiva (TAI)

Además de la TABR, la TAE y la TAN, se ha desarrollado recientemente una nueva funcionalidad que se puede sumar a estos paradigmas de TA. Es la traducción automática interactiva (TAI) o adaptativa (Green, 2016). El objetivo de esta nueva funcionalidad es aumentar la productividad y la calidad de traducción aún más. La TAI se propuso originalmente en el proyecto TransType (Langlais et al., 2000), aunque se mejoró posteriormente en el proyecto TransType 2 (Tomás y Casacuberta, 2006). Pese

a que hace 20 años de su concepción, aún se conoce muy poco y no se le ha dado tanta importancia como a la TAE y la TAN.

Este paradigma de interacción persona-máquina requiere de la participación tanto de la persona como del sistema informático. Mediante un proceso iterativo que funciona segmento a segmento, el sistema sugiere una primera propuesta con una traducción en bruto y, a continuación, la persona encargada de la posedición modifica lo que considere que no es correcto. El sistema de TAI —sea en un motor de TAE o TAN— tiene en cuenta esta modificación y sugiere una nueva traducción en bruto, conservando lo que el poseditor ya ha aceptado (Barrachina et al., 2009).

Source (x):		They are lost forever .
Target (ŷ):		Ils sont perdus à jamais .
IT-0	**MT**	Ils sont perdus pour toujours .
IT-1	**User**	*Ils sont perdus* **à** pour toujours .
	MT	*Ils sont perdus à* jamais .
IT-2	**User**	*Ils sont perdus à jamais .*

Figura 1. Proceso de un motor de TAI. Imagen extraída de «Active Learning for Interactive Neural Machine Translation of Data Streams» (Peris y Casacuberta, 2018).

En la Figura 1 (Peris y Casacuberta, 2018) podemos ver el proceso de un motor de TAI en la combinación lingüística inglés-francés. «They are lost forever» es una oración que funciona como un ejemplo y que hace la función de texto origen. «Ils sont perdus à jamais» es la traducción correcta del TO y funciona como texto de destino. «IT-» hace referencia a cada paso iterativo que hace el motor. «MT» se refiere a la traducción en bruto que el motor de TAI ofrece en cada proceso iterativo y «User» muestra las modificaciones que introduce la persona poseditora. El elemento que aparece dentro de la caja negra es la modificación, adición u omisión exacta de la traducción en bruto ofrecida por el motor, y las palabras que aparecen en verde son las que ya ha validado la persona que posedita.

Para explicar este proceso más concretamente, cuando la oración en inglés «They are lost forever» se introduce en el motor de TAI, este sugiere «Ils sont perdus pour toujours» en francés. Esta traducción no sería adecuada para el texto origen introducido. A continuación, podemos ver que, en IT-1, el poseditor (o *User*) acepta «Ils sont perdus» como una traducción correcta. A continuación, introduce «à» (la palabra que aparece dentro de la caja). Es en este momento cuando el motor de TAI acepta la validación humana y tiene en cuenta la adición, que vuelve a ejecutar su sistema y ofrece una nueva traducción en bruto: «Ils sont perdus à jamais».

Dicho esto, la TAI funciona de una manera similar a los teclados y textos predictivos de un teléfono móvil o que herramientas como Gmail están empezando a introducir. El tiempo de procesamiento es un factor clave a la hora de volver a ejecutar estos sistemas y, con los algoritmos y la potencia de procesamiento creciente de los ordenadores actuales, esta restricción se reduce y la TAI se vuelve cada vez más atractiva (Bender et al., 2005; Koehn y Haddow, 2009; Peris y Casacuberta, 2019).

3.3. La calidad de traducción

Una vez conocemos los modelos y los sistemas de TA disponibles en la actualidad y de vanguardia, toca abordar uno de los temas más importantes y complejos en los estudios de tecnologías de la traducción actuales: la «calidad».

La situación descrita en los apartados anteriores y la globalización han transformado el panorama y la industria de la traducción enormemente. Actualmente, hay una cantidad ingente de contenido por traducir y, además, los clientes cada vez exigen plazos de entrega menores y con presupuestos más bajos. No obstante, quien necesita una traducción y la encarga a un proveedor de servicios lingüísticos quiere asegurarse de que esta es correcta. Es aquí, y en cualquier otra necesidad de traducción, cuando se habla de «calidad». En el volumen 12 de la *Revista Tradumática*, que versa sobre traducción y calidad, hay tres artículos de varios autores que debaten el término calidad (Fields et al., 2014; Koby et al., 2014; A. Melby et al., 2014). Aun así, los autores

discrepan a la hora de ofrecer una definición final. En primer lugar, ofrecen una definición amplia:

> A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs. (Koby et al., 2014)

No obstante, también afirman que no acaban de estar convencidos de esta definición porque podría incluir otros servicios lingüísticos, como por ejemplo la transcreación o la localización. Por tanto, ofrecen una segunda definición más concreta:

> A high-quality translation is one in which the message embodied in the source text is transferred completely into the target text, including denotation, connotation, nuance, and style, and the target text is written in the target language using correct grammar and Word order, to produce a culturally appropriate text that, in most cases, reads as if originally written by a native speaker of the target language for readers in the target culture. (Koby et al.,2014)

Así pues, las dos definiciones mencionan a las mismas partes clave implicadas en el proceso de traducción: i) un texto en LO que ii) se transmite completamente, con fluideza y adecuación, a iii) otro texto en LM que iv) cumple los requisitos del usuario final y de la cultura meta. Expuesto esto, queda claro que el término calidad es complejo, ya que puede variar en función de qué prefiere el cliente, si una traducción de asimilación o de diseminación (Ginestí y Forcada, 2009). Sin embargo, es un término que está adoptado ampliamente en el mundo de la traducción (Sánchez-Gijón, 2014).

Además, según Leiva (2018, pp. 261-264), los proveedores de servicios lingüísticos están empezando a utilizar cada vez más una serie de certificaciones y acreditaciones que varios organismos otorgan tras hacer un seguimiento de los procesos de traducción (por ejemplo, la viabilidad, la gestión, la entrega y que las traducciones pasen por un traductor y un revisor cualificados y certificados). Algunas de las más destacadas son las normas ISO 17100:2015 (ISO, 2015), la ISO 18787:2018 (ISO, 2017) y la ISO 9001:2015

(ISO, 2015a). Pese a que el hecho de estar en posesión de una de estas acreditaciones puede indicar que se siguen ciertos procesos para asegurar la calidad, no se hace ninguna evaluación de los textos para asegurarlo. Estas solo se basan en procesos.

3.4. La evaluación de la calidad

Los proveedores de servicios lingüísticos, los usuarios de motores de traducción o incluso los desarrolladores de estos últimos necesitan corroborar que las traducciones cumplen con unos criterios mínimos de calidad o que las traducciones que ofrecen son correctas. Para conseguir esto, aparece la evaluación de la calidad, que ha recibido mucha atención en los últimos años. A causa de la gran difusión reciente de la TA y la gran cantidad de dinero que se está invirtiendo, ha causado que haya dos eventos internacionales de gran calado al respecto. En primer lugar, el «Workshop on Statistical Machine Translation (WMT)» (véase http://www.statmt.org/wmt20/), que se celebra de manera anual desde el 2006 (y desde hace años incluye la TAN). Este primer evento se centra en idiomas europeos y tiene dos circuitos: uno de traducción y otro de evaluación. Además, tiene dos direcciones diferenciadas: del inglés a otros idiomas y viceversa. En segundo lugar, está el «International Workshop of Spoken Language Translation (IWSLT)» (véase http://iwslt.org/doku.php), el cual se centra en la traducción oral y los idiomas asiáticos. Asimismo, hay dos formas de evaluación ampliamente aceptadas por la academia y la industria: la evaluación automática y la evaluación humana.

3.4.1. La evaluación automática de la calidad

El interés en la TA reside en dos únicos principios: el aumento de la productividad y la reducción de los costes. Tal y como se ha mencionado antes, actualmente hay mucho material que traducir y se quiere hacer más rápido, así como también de manera más barata. Al crear un motor de TA, el objetivo debe ser poder traducir más en menos. Para saber si esto será posible, los desarrolladores y los informáticos que se encargan de estas tareas necesitan corroborar y comprobar si su motor rinde bien o no, así como si las modificaciones que se han introducido han servido para mejorar e ir hacia adelante, en lugar de hacia atrás. Y esto debe hacerse de una manera sencilla y rápida, que

permita seguir el ritmo al que avanzan la sociedad, la industria y el mercado. Según (Martín-Mor et al., 2016, pp. 63-64):

> Actualment, la recerca en avaluació de qualitat de TA se centra en l'afinació d'índexs de qualitat per mitjà de la comparació de traduccions en brut amb traduccions humanes de referencia (també conegudes com a golden standard), o bé amb un corpus comparable de textos en la llengua d'arribada. Si existeix una traducció humana del mateix text, es compara cada segment en termes de nombre d'edicions (insercions, eliminacions i substitucions) necessari per convertir cada segment de la traducció en brut en el segment de la traducció humana.

A pesar de que han ido surgiendo muchas métricas automáticas de evaluación de la calidad, solo haré un repaso y una breve explicación de las más utilizadas y destacadas.

En primer lugar, aparece una métrica que Ye-Yi Wang et al. (2003) nombran WER (Word Error Rate). En esta métrica, se utiliza una traducción humana como referencia y «Golden Standard», con el objetivo de calcular el número mínimo de ediciones necesarias para que la traducción en bruto de un motor de TA se corresponda con la oración de referencia. Además, tiene en cuenta las inserciones, eliminaciones y sustituciones de palabras en la oración, así como el orden de las palabras con el objetivo de hacer coincidir dos oraciones con el número de cambios más pequeño posible.

Según Han (2018), más tarde se comprueba que el WER tiene varios problemas: como, por ejemplo, no calcula bien el orden de las palabras y no tiene en cuenta las coincidencias aproximadas como los sinónimos, ya que es una métrica automática que no puede tener en cuenta las particularidades y complejidades del lenguaje completamente. Como consecuencia, se proponen otras métricas para solventar este problema: el PER (Tillmann et al., 1997) y el TER (M. Snover et al., 2016). El TER (*Translation Edit Rate*) se convierte en una métrica bastante utilizada en la academia y la industria, la cual es una leve mejora de las dos anteriores, ya que añade como posible edición el cambio de una palabra o grupo de palabras de una oración a otra parte de la

oración. No obstante, han continuado publicándose mejoras continuamente, por ejemplo, el TER-Plus (Snover et al., 2009).

Estas métricas automáticas anteriores miden la distancia de edición (es decir, cuántos cambios son necesarios para homogeneizar el texto en bruto de la TA y una *Golden standard*). Por otra parte, hay un tipo de métricas automáticas que se centran en un aspecto diferente: el orden de las palabras o de los grupos de palabras.

En este grupo de métricas aparece el BLEU (Papineni et al., 2001). El BLEU (*Bilingual Evaluation Understudy*), según sus autores, funciona de la siguiente manera: «to compare n-grams of the candidate with the n-grams of the reference translation and count the number of matches. These matches are position-independent. The more the matches, the better the candidate translation is». En otras palabras, esta métrica calcula la frecuencia con la que las palabras o frases (hasta un conjunto de 4 palabras) coinciden tanto en la referencia humana y la traducción en bruto del motor de TA. Uno de los problemas de esta métrica es que utiliza una oración humana como referencia. No obstante, una misma oración en la LO puede traducirse de múltiples maneras en la LM de forma correcta. Por tanto, es posible que se valore negativamente una oración o una traducción correcta. En consecuencia, cuantas más referencias humanas se proporcionen, mejor funciona BLEU.

Al igual que en el grupo de métricas automáticas anteriores, posteriormente se proponen algunas modificaciones del BLEU que cambian la ponderación y la evaluación de la TA. Aparece NIST (Doddington, 2002), que suma un peso especial a las palabras «raras» o poco utilizadas, que salían bastante perjudicadas al utilizar BLEU. Asimismo, también se presenta METEOR (Banerjee y Lavie, 2005), una métrica adicional que cambia ligeramente la evaluación. En lugar de calcular los n-gramas, según sus autores, METEOR «is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations. Unigrams can be matched based on their surface forms, stemmed forms, and meanings».

Dicho esto, es muy importante destacar que las métricas automáticas de evaluación de la calidad tienen algunos problemas que se han comprobado ampliamente. A menudo, BLEU se utiliza en evaluaciones segmento a segmento, y esto perjudica la evaluación, ya que es un método diseñado para hacer evaluaciones en el nivel del documento (Way, 2018). Otro problema es que, como las métricas son estadísticas, pueden influir en la evaluación de motores diferentes (por ejemplo, TABR y TAE) y evaluar mejor una traducción hecha por un motor de TAE. Además, He y Way (2009) también indican que METEOR prefiere traducciones automáticas más largas y las valora mejor, mientras que TER las prefiere más cortas. A pesar de esto y, tal como veremos a continuación, la evaluación humana es muy cara, no puede reproducirse y consume mucho tiempo. Como consecuencia, es necesario que los desarrolladores de motores de TA puedan acceder a una serie de evaluaciones rápidas, baratas y automáticas, que les facilite saber ligeramente si el motor ha mejorado o empeorado. Dicho esto, BLEU se ha consolidado como métrica por excelencia y es la que se utiliza como marco de referencia en la mayoría de los estudios sobre evaluación de calidad de TA.

3.4.2. La evaluación humana de la calidad

De conformidad con lo explicado en el punto anterior, las métricas automáticas de evaluación de la calidad de TA tienen algunos puntos flojos, ya que el lenguaje entraña obstáculos, complejidades y dificultades que no son fáciles de representar o calcular con métricas automáticas. Además, el aumento de la calidad de los nuevos motores de TA está creando nuevos retos y dificultades a la hora de evaluar la calidad. Aunque la evaluación y las métricas automáticas son útiles para los desarrolladores de motores de TA, Läubli et al. (2020) indican que «human evaluation remains the gold standard, but there are many design decisions that potentially affect the validity of such a human evaluation» y que «there is consensus that a reliable evaluation should (despite high costs) be carried out by humans». Por tanto, diseñar una buena metodología para evaluar la TA ha sido uno de los objetivos principales en el sector de las tecnologías de la traducción en las últimas décadas. Se han propuesto múltiples métodos de evaluación humana de TA (Moorkens et al., 2018). En el capítulo de evaluación de TA de la *Routledge Encyclopedia for Machine Translation* se proponen varios métodos de manera global y general: puntuar los segmentos del 1 al 5 de conformidad con la calidad;

hacer una clasificación relativa; analizar los errores mediante métricas MQM-DQF (Görög, 2014; O'Brien, 2012); mediante pruebas de comprensión; o posedición. No obstante, en los eventos internacionales de gran calado como el WMT y el IWSLT se han utilizado principalmente dos métodos:

- Clasificación relativa: en este primer método de evaluación, las personas que evalúan la traducción obtienen la oración original en LO y varias opciones de motores de TA en la LM. De esta manera, las personas evaluadoras asignan un orden de calidad y clasifican de manera relativa qué motores funcionan mejor. Por ejemplo, el motor A es el mejor, el motor B es el segundo mejor y el motor C es el peor (Bojar et al., 2016; Koehn y Monz, 2006).
- Evaluación directa: en este segundo método de evaluación, las personas evaluadoras obtienen la oración original en la LO y van viendo las oraciones traducidas (ya sean diferentes motores de TA o traducciones humanas) de uno en uno. En este momento, tienen que asignar una puntuación del 0 al 100 a cada traducción. Este método permite saber qué motor es mejor y, además, también permite saber en qué grado lo es. Asimismo, para evitar problemas de subjetividad y de puntuación entre los diferentes evaluadores, previamente se preparan documentos con instrucciones para homogeneizar los criterios (Graham et al., 2013).

Según Callison-Burch et al. (2007), la evaluación por clasificación relativa proporciona un acuerdo mayor entre las personas evaluadoras que los otros métodos, permitiendo saber que el motor de A de TA es mejor que el B y el C. No obstante, no se puede saber si es mucho mejor o solo un poco, ya que la evaluación es relativa, tal y como se ha mencionado.

3.5. La posedición

Habida cuenta de la situación actual de las tecnologías de la traducción y de la TA, con sus modelos de vanguardia y calidad actual, es muy importante indicar cuál es el uso principal que se le da. Recientemente, algunos estudios han indicado y afirmado que

algunos sistemas de TAN de vanguardia ofrecen traducciones que están al nivel de los traductores profesionales, o incluso que les superan (Bojar et al., 2018; Hassan et al., 2018; Popel, 2018; Wu et al., 2016). No obstante, si la calidad quiere asegurarse, continúa siendo necesaria e imprescindible la intervención humana (Castilho et al., 2017; Läubli et al., 2020). Esta intervención tiene lugar a través de la posedición. Según (O'Brien, 2011, pp. 197-198), la posedición es «the correction of raw machine translated output by a human translator according to specific guidelines and quality criteria». Esta tarea sigue el siguiente proceso:

1. un texto digital en la LO se envía a un motor de TA;
2. este motor ofrece una propuesta de texto digital en la LM; y
3. la persona poseditora modifica, cambia o altera el texto en bruto según sea necesario.

De esta manera, la posedición se utiliza con el objetivo de aumentar la productividad de los traductores, ya que, gracias a la TA, se puede traducir más contenido en menos tiempo. En los últimos años, la posedición ha sido objeto de un gran estudio en el campo de las tecnologías de la traducción, y se ha investigado desde muchos ámbitos, ángulos y puntos de vista diferentes: se ha comparado el esfuerzo de posedición con otros tipos de traducción asistida por ordenador (O'Brien y Moorkens, 2014); se ha medido la productividad y la calidad de posedición en comparación con otros tipos de traducción tradicional (Guerberof-Arenas, 2008); se han investigado las necesidades de las personas poseditoras al poseditar (Moorkens y O'Brien, 2017); se ha medido la satisfacción de las personas poseditoras (Cadwell et al., 2016); y se ha comparado la productividad al poseditar TAE y TAN (Läubli et al., 2018; Sánchez-Gijón et al., 2019), por nombrar algunos estudios destacados de la materia.

4. Metodología

Una vez expuesta la situación actual de las tecnologías de la traducción, la TA y la evaluación de la TA, es el momento de definir y dar forma al marco metodológico para conseguir los objetivos del punto 2. Justificación y objetivos. La metodología del

presente estudio se ha basado en un trabajo similar para una evaluación humana en la combinación inglés-español (López-Pereira, 2019). Para conseguir el primer objetivo, se ha hecho una revisión extensa de la bibliografía más actualizada sobre la TA y la evaluación de TA (véase el apartado 3. Contextualización del objeto de estudio). En cuanto al segundo y tercer objetivo, el presente trabajo evaluará y versará sobre tres motores de TA diferentes en la combinación de idiomas inglés-catalán: Apertium (un motor de TABR de código libre), el Traductor de Softcatalà (un motor de TAN de código libre) y el Traductor de Google (un motor de TAN de código privativo). Se proporcionará más información de los mismos en el apartado 4.1.1. Motores.

En la presente investigación, hay dos objetivos principales (objetivos 2 y 3). En primer lugar, saber qué motor ofrece una calidad de traducción mayor. Aquí se parte de la hipótesis que los motores de TAN ofrecerán una calidad de traducción mayor (en términos de precisión y fluidez) que el motor de TABR. No obstante, se desconoce cuál de los dos motores de TAN (el de Softcatalà o el de Google) ofrecerá una calidad mayor. En segundo lugar, se pretende saber qué sistema ofrece un rendimiento mayor al poseditar, para así descubrir si interesa más introducir el motor de TABR o uno de los dos de TAN en un flujo de trabajo cuyo objetivo sea la diseminación. Pese a que la TABR pueda producir errores más claros, estos siempre se repetirán y puede ser más fácil detectarlos y editarlos. No obstante, algunos estudios indican que la TAN consigue oraciones fluidas y gramaticalmente correctas, pese a que puede tener errores de significado. Por esta misma razón, tal vez la TAN sea un arma de doble filo y, a la hora de poseditar, requiera de más tiempo para encontrar los errores subyacentes y conseguir una traducción con calidad humana.

Como se ha establecido anteriormente, las evaluaciones automáticas no son certeras a la hora de determinar la calidad de la TA. En consecuencia, para conseguir los objetivos del presente trabajo, se hará una evaluación humana que constará de cuatro partes bien diferenciadas: i) una evaluación de clasificación relativa (la prueba *MT Ranking* de TAUS); ii) una evaluación de la fluidez (la prueba *Fluency* de TAUS); iii) una evaluación de la precisión (la prueba *Accuracy* de TAUS); y una evaluación de la productividad (la prueba *Post-editing Productivity* de TAUS). La segunda y tercera prueba (fluidez y

precisión) se hacen simultáneamente, en una sola interfaz. Todas estas evaluaciones se realizarán con el objetivo de evaluar la calidad (clasificación relativa, fluidez y precisión) y la productividad (a través de un análisis de la posedición). Las variables del estudio y los experimentos se explican a continuación.

4.1. Variables del estudio

En este apartado se presentarán los elementos clave e indispensables para llevar a cabo el experimento, a saber: se explicarán brevemente los motores que se utilizarán; se definirá qué texto se ha elegido y por qué motivos se ha seleccionado dicha tipología específica; y se expondrán los participantes que participarán en el estudio, tanto para la evaluación de la calidad, como para la evaluación de la productividad.

4.1.1. Motores

Tal y como se ha mencionado anteriormente, el presente trabajo de investigación evaluará y tratará tres motores diferentes:

- Apertium: es un motor de TABR que nace tras la financiación pública a un proyecto de investigación, originalmente destinado a idiomas cercanos (castellano-catalán), pero que posteriormente se ha expandido para ofrecer servicios de TA a otras combinaciones lingüísticas que incluyen el inglés, el catalán, el valenciano, el aragonés, el sardo o el asturiano, entre otros muchos idiomas. Lo interesante de este motor o plataforma es que es de código libre y, por lo tanto, i) es gratuito, ii) puede ejecutarse libremente para cualquier fin, iii) puede distribuirse libremente y iv) el código fuente está disponible en línea y puede modificarse o utilizarse libremente (https://www.apertium.org/).
- Traductor de Softcatalà: es un motor de TAN de código libre que se ha publicado recientemente y que ha creado y entrenado el equipo de Softcatalà. Softcatalà es una asociación sin ánimo de lucro que tiene el objetivo de fomentar el uso del catalán en la informática, Internet y las nuevas tecnologías. El motor está entrenado con corpus de traducciones de productos digitales de código libre y abierto (entre los que se encuentran proyectos de Firefox, Linux, Wikimatrix,

etc.), el Diario Oficial de la Generalitat de Catalunya y de la Generalitat Valenciana, así como documentos con la licencia Creative Commons (https://www.softcatala.org/traductor-angles-catala/).

- Traductor de Google: es un motor de TA de código privativo desarrollado por Google. Utiliza TAN desde inicios del 2020. Es tal vez uno de los traductores automáticos más utilizados mundialmente, pero al ser de código privativo conservan las traducciones y puede haber conflictos de confidencialidad. Además, un aspecto importante que hay que describir a la hora de hablar del Traductor de Google son las lenguas pivote. En algunas combinaciones lingüísticas donde la presencia de textos paralelos multilingües con traducciones de calidad no abunda, se suele hacer primero una traducción entre una combinación con grandes corpus disponibles y, a continuación, se hace una segunda traducción desde una lengua pivote. En la versión antigua, el motor TAE, sí que se sabe que utilizaban el castellano como lengua pivote para traducir del inglés al catalán. No obstante, al ser un motor de TA privativo, no se sabe si en el paso de TAE a TAN y si en la combinación de inglés-catalán continúan utilizando estas lenguas intermedias (https://translate.google.es/#view=home&op=translate&sl=en&tl=ca).

En este momento, es importante destacar dos conceptos clave con respecto a los motores: «in-domain» y «out-of-domain». Los traductores automáticos neuronales se entrenan con grandes cantidades de corpus lingüísticos. Es de especial relevancia el tipo de texto, ya que, si un motor está entrenado con una gran cantidad de textos de un campo o especialidad concreta, parece indicar que ofrecerá una calidad mayor en dicho campo o especialidad («in-domain») que en un texto de una especialidad que el motor no haya visto en el entrenamiento («out-of-domain») (Doğru et al., 2019; Etchegoyhen et al., 2018). Por tanto, es importante destacar que el motor de Softcatalà ha utilizado una gran cantidad de memorias de traducción de productos digitales de software libre a la hora de entrenar el motor. En cambio, para el Traductor de Google, no conocemos qué corpus se han utilizado en el entrenamiento.

4.1.2. Selección y obtención de los textos

Un aspecto adicional muy importante es la elección del texto que se utilizará en el presente estudio. En este estudio, se utilizarán fragmentos de Home Assistant (véase https://www.home-assistant.io/), un asistente virtual para casas inteligentes. Los motivos de elección de este texto es que, al ser un producto de código libre, se pueden obtener fácilmente los segmentos y el texto para trabajar. Además, la localización de software es uno de los tipos de traducción más habituales en el mundo profesional actual y el mundo del voluntariado y el activismo lingüístico (especialmente, en lenguas minorizadas o sin estado). En consiguiente, el objetivo del trabajo es obtener unos resultados aptos y válidos para la situación actual. La selección del texto se ha realizado de manera coordinada con Softcatalà.

Para descargar el archivo con el texto que se utilizará como muestra en la presente evaluación, hay que entrar al repositorio de Github de Home Assistant y encontrar el archivo «homeassistant.json», un formato de archivo que se utiliza bastante en el mundo de la localización. Además, en este archivo podremos encontrar todas las cadenas traducibles de Home Assistant.

4.1.3. Preparación del texto

Tras encontrar el archivo JSON, se utilizarán las herramientas de Okapi Framework para convertir el archivo JSON a un archivo TXT y así poder eliminar todas las etiquetas de código y poder trabajar de manera más sencilla.

En este caso, el proyecto Home Assistant incluye 1 766 segmentos, una cantidad demasiado grande como para hacer una evaluación humana. Por tanto, tendremos que tratar el archivo para poder extraer el texto que nos interesa de la LO y poder preparar los archivos para la evaluación, de conformidad con los archivos requeridos por las herramientas TAUS DQF que se han explicado anteriormente.

En primer lugar, se ha hecho una selección manual y aleatoria de los segmentos. Se han seleccionado 100 segmentos para cada texto (Texto 1 y Texto 2), y el número de palabras resultante ha sido el siguiente: 1 006 palabras (UI Sample 1) y 1 065 palabras

(UI Sample 2). En esta selección, se han mezclado segmentos largos y con segmentos cortos (ya que estos últimos son muy frecuentes en localización de software), así como también con *placeholders* y variables.

Los textos se pueden observar en la hoja de datos «HomeAssistant – Completo» del Anexo 4. En la primera pestaña están todos los segmentos del texto original, 1 766 segmentos, y en la segunda y tercera pestaña están los segmentos escogidos con sus respectivas traducciones automáticas. Los segmentos del Texto 1 se han marcado con el color amarillo, tanto en los segmentos en la LO como en la segunda pestaña; y los segmentos del Texto 2 se han marcado de azul, tanto en los segmentos en la LO como en la tercera pestaña.

Una vez se han seleccionado ambos textos, se ha procedido a traducir los segmentos con los motores de TA de manera manual. Obtenidas las traducciones en bruto tanto del Traductor de Softcatalà, como del Traductor de Google y de Apertium, se han volcado en la hoja de datos «HomeAssistant – Completo» (véase el Anexo 4. Hojas de datos del estudio). Para poder conformar las evaluaciones que necesitamos para el estudio, se han preparado hojas de Excel adicionales según el tipo de evaluación y de conformidad con los requisitos de las plataformas DQF de TAUS. Estos requisitos se explicarán en los apartados 4.2.1 MT Ranking y 4.2.4 Productividad siguientes. Tras preparar los archivos necesarios, se suben a la herramienta DQF de TAUS y se introduce los correos electrónicos de las personas evaluadoras. De esta manera, reciben un correo electrónico automático con un enlace que les dirige a la evaluación correspondiente (sea MT Ranking, Precisión y Fluidez o Productividad).

4.1.4. Personas evaluadoras

Según Läubli et al. (2020), la selección de las personas que harán la evaluación también es muy importante a la hora de definir la metodología. La evaluación humana la pueden llevar a cabo tanto traductores profesionales como aficionados. Es más, Moorkens et al. (2018, p. 23) indican que hay una tendencia a «rely on students and amateur evaluators, sometimes with an undefined (or self-rated) proficiency in the languages involved, an

unknown expertise with the text type», ya que es más fácil conseguir su colaboración. Asimismo, Castilho et al. (2017) también destacan lo siguiente:

> Non-expert translators lack knowledge of translation and so might not notice subtle differences that make one translation more suitable than another, and therefore, when confronted with a translation that is hard to post-edit, tend to accept the MT rather than try to improve it.

En consecuencia, se ha tenido muy en cuenta la selección de las personas evaluadoras. Uno de los objetivos rectores ha sido que tuvieran conocimientos similares, ya que, si no, los resultados obtenidos podrían variar mucho (según la especialidad y la experiencia de cada persona). Por lo tanto, para intentar reducir la diferencia y buscar un grupo homogéneo de participantes, se han elegido las siguientes personas:

- Para la prueba MT Ranking, han participado en la evaluación 11 personas. Todas tienen un grado en Traducción e Interpretación, un máster en Traducción Especializada (sea en el ámbito tecnológico o en otro ámbito de especialidad) y tienen el catalán como lengua materna. Además, todas las personas tienen de 1 a 3 años de experiencia profesional en la industria de la traducción, ya sea como traductores, revisores, poseditores o gestores de proyectos. Se ha elegido este perfil, ya que era un tipo de sujetos con formación en traducción, con conocimientos técnicos y suficientes para poder evaluar una traducción especializada como la localización de software y con algo de experiencia profesional.
- Para la prueba de Precisión y Fluidez, la evaluación la ha hecho el autor, que está graduado en Traducción e Interpretación, tiene un máster en Traducción Especializada (en el ámbito de las tecnologías de la traducción) y 3 años de experiencia profesional como traductor autónomo y traductor jurado. Además, cuenta con 2 años de experiencia profesional en localización de software.
- Para la prueba de evaluación de la productividad se han creado dos grupos de estudio:

- El primer grupo de estudio ha evaluado el Traductor de Softcatalà y el Traductor de Google. El grupo está compuesto por 6 personas que han participado también en la primera prueba, la evaluación MT Ranking.
- El segundo grupo de estudio ha evaluado el Traductor de Softcatalà y Apertium. El grupo está compuesto por 6 miembros de Softcatalà, con experiencia en traducción y localización de productos digitales como voluntarios, pero que no necesariamente han estudiado Traducción e Interpretación ni se dedican a traducir profesionalmente.

La selección de participantes se ha hecho de esta manera por varios motivos, que aparecen en las descripciones de las pruebas a continuación (véase 4.2.1. MT Ranking, 4.2.2. Fluidez, 4.2.3. Precisión y 4.2.4. Productividad).

4.2. Evaluación humana

A la hora de hacer la evaluación humana, se utilizarán las herramientas Dynamic Quality Framework de TAUS (véase https://dqf.taus.net/). Estas herramientas están desarrolladas por TAUS, una organización de la industria de la traducción que investiga y ofrece servicios en tecnologías de la traducción y TA.

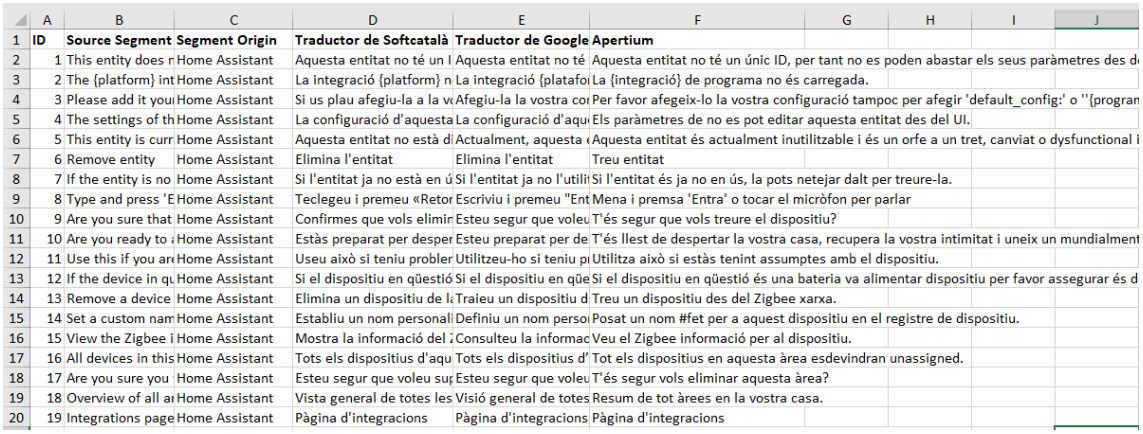
Es en este momento en el que podrá denotarse uno de los problemas de la evaluación humana: la subjetividad de las personas evaluadoras. Para evitar este factor o, dentro de lo posible, reducirlo, se diseñarán unas guías de evaluación que deberán leerse detenidamente antes de realizar la diferentes pruebas y que servirán para homogeneizar los criterios. Estas guías pueden encontrarse en el Anexo 1. Pautas para la evaluación «MT Ranking», el Anexo 2. Pautas para la evaluación «Precisión y fluidez» y el Anexo 3. Pautas para la evaluación «Productividad». Además, se pueden ver capturas de pantalla de la interfaz donde se hará la evaluación, para que así las personas evaluadoras puedan conocer de antemano la plataforma DQF de TAUS. Con arreglo a lo mencionado en el apartado 4. Metodología anterior, las diferentes pruebas de la evaluación humana se llevarán a cabo en el siguiente orden.

4.2.1. MT Ranking

En esta primera prueba, y de conformidad con uno de los modos explicados en el apartado 3.4.2. La evaluación humana de la calidad, utilizaré una prueba de clasificación relativa. Así pues, se introducirán los resultados de traducción de los tres motores de TA escogidos. En esta primera prueba, habrá 11 personas evaluadoras, todas con un perfil similar: grado en Traducción e Interpretación, máster en Traducción (ya sea de perfil tecnológico o de especialización) y con una experiencia profesional de 1 a 3 años en el sector (como traductores autónomos, en plantilla o en el departamento de gestión de proyectos).

Las personas evaluadoras podrán ver una oración en la LO y, a continuación, verán tres propuestas de traducción. Las propuestas corresponderán a cada uno de los motores (Apertium, Traductor de Softcatalà o Traductor de Google) y serán anónimas; las personas evaluadoras no conocerán a qué motor corresponde cada una. De esta manera, deberán clasificar relativamente qué traducciones son las mejores. Por ejemplo: la opción A es la mejor (1), la opción B es la segunda mejor (2) y la opción C es la peor (3). En el caso de que la opción A y B sean igual de buenas, ambas recibirían la puntuación 1 y se asignaría la puntuación 3 a la C.

Para esta primera prueba, la plataforma DQF de TAUS solo acepta archivos XLS, XLSX, CSV o TMX. Además, los archivos deben tener varias columnas obligatorias (ID del segmento, segmento en la LO, nombre del archivo en la LO, TA del motor 1 y TA del motor 2; en caso de que se comparen tres motores, también la TA del motor 3). Véase la imagen a continuación:



	A	B	C	D	E	F	G	H	I	J
1	ID	Source Segment	Segment Origin	Traductor de Softcatalà	Traductor de Google	Apertium				
2	1	This entity does n	Home Assistant	Aquesta entitat no té un I	Aquesta entitat no té	Aquesta entitat no té un únic ID, per tant no es poden abastar els seus paràmetres des d				
3	2	The {platform} int	Home Assistant	La integració {platform} n	La integració {platafo	La {integració} de programa no és carregada.				
4	3	Please add it you	Home Assistant	Si us plau afegiu-la a la v	Afegiu-la la vostra co	Per favor afegeix-lo la vostra configuració tampoc per afegir 'default_config:' o '{progran				
5	4	The settings of th	Home Assistant	La configuració d'aquesta	La configuració d'aqu	Els paràmetres de no es pot editar aquesta entitat des del UI.				
6	5	This entity is curr	Home Assistant	Aquesta entitat no està d	Actualment, aquesta	Aquesta entitat és actualment inutilitzable i és un orfe a un tret, canviat o dysfunctional i				
7	6	Remove entity	Home Assistant	Elimina l'entitat	Elimina l'entitat	Treu entitat				
8	7	If the entity is no	Home Assistant	Si l'entitat ja no està en ú	Si l'entitat ja no l'utilit	Si l'entitat és ja no en ús, la pots netejar dalt per treure-la.				
9	8	Type and press 'E	Home Assistant	Teclegeu i premeu «Retor	Escriviu i premeu "Ent	Mena i premsa 'Entra' o tocar el micròfon per parlar				
10	9	Are you sure that	Home Assistant	Confirmes que vols elimir	Esteu segur que voleu	T'és segur que vols treure el dispositiu?				
11	10	Are you ready to (	Home Assistant	Estàs preparat per desper	Esteu preparat per de	T'és llest de despertar la vostra casa, recupera la vostra intimitat i uneix un mundialment				
12	11	Use this if you ar	Home Assistant	Useu això si teniu probler	Utilitzeu-ho si teniu p	Utilitza això si estàs tenint assumptes amb el dispositiu.				
13	12	If the device in qu	Home Assistant	Si el dispositiu en qüestió	Si el dispositiu en qüe	Si el dispositiu en qüestió és una bateria va alimentar dispositiu per favor assegurar és d				
14	13	Remove a device	Home Assistant	Elimina un dispositiu de la	Traieu un dispositiu d	Treu un dispositiu des del Zigbee xarxa.				
15	14	Set a custom nam	Home Assistant	Estableix un nom personal	Definiu un nom perso	Posat un nom #fet per a aquest dispositiu en el registre de dispositiu.				
16	15	View the Zigbee i	Home Assistant	Mostra la informació del z	Consulteu la informac	Veu el Zigbee informació per al dispositiu.				
17	16	All devices in this	Home Assistant	Tots els dispositius d'aqu	Tots els dispositius d'	Tot els dispositius en aquesta àrea esdevindran unassigned.				
18	17	Are you sure you	Home Assistant	Esteu segur que voleu su	Esteu segur que voleu	T'és segur vols eliminar aquesta àrea?				
19	18	Overview of all ar	Home Assistant	Vista general de totes les	Visió general de totes	Resum de tot àrees en la vostra casa.				
20	19	Integrations page	Home Assistant	Pàgina d'integracions	Pàgina d'integracions	Pàgina d'integracions				

Figura 2. MT Ranking: Ejemplo de un archivo preparado para la plataforma DQF de TAUS

4.2.2. Fluidez

En esta segunda prueba de evaluación se pretende evaluar la fluidez de las traducciones automáticas. Según Linguistic Data Consortium (véase https://www.ldc.upenn.edu/), la fluidez es el grado en que una traducción «is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker». Es decir, hay que indicar el grado en que las traducciones son correctas gramaticalmente, son naturales en la LM y se adaptan a sus normas lingüísticas y ortográficas. Según DQF de TAUS, debe puntuarse el grado de fluidez de las oraciones de la siguiente manera:

- 1. Incomprehensible (Incomprensible): «refers to a very poorly written text that is impossible to understand».
- 2. Disfluent (Mala fluidez): «refers to a text that is poorly written and difficult to understand».
- 3. Good (Buena fluidez): «refers to a smoothly flowing text even when a number of minor errors are present».
- 4. Flawless (Perfecta): «refers to a perfectly flowing text with no errors».

4.2.3. Precisión

En esta tercera prueba, se pretende evaluar la precisión de las traducciones automáticas. Según Linguistic Data Consortium, la precisión es «how much of the meaning expressed in the gold-standard translation or the source is also expressed in

the target translation» (véase https://www.ldc.upenn.edu/). Es decir, se deberá indicar el grado en que la traducción propuesta por un motor de TA (anónimo) expresa y recoge el significado de la oración en la LO. Según DQF de TAUS, se debe puntuar el grado de precisión de las oraciones de la siguiente manera:

- 1. None (Nada): «none of the meaning in the source is contained in the translation».
- 2. Little (Ligeramente): «fragments of the meaning in the source are contained in the translation».
- 3. Most (Mayoritariamente): «almost all the meaning in the source is contained in the translation».
- 4. Everything (Completamente): «all the meaning in the source is contained in the translation, no more, no les».

La evaluación de la segunda y la tercera prueba las hará el autor del estudio, personalmente. En el mejor de lo casos, hubiera sido ideal conseguir personas adicionales que evaluaran esta segunda y tercera prueba y que no fuera el propio autor. No obstante, por las limitaciones de tiempo a la hora de hacer el presente estudio, esto no ha sido posible. Como el autor no ha participado en las otras evaluaciones y, con el fin de obtener unos resultados homogéneos y más fáciles de analizar, será quien haga la prueba de precisión y fluidez. Como recurso adicional para homogeneizar los criterios en la evaluación de las pruebas —independientemente de ser el investigador la persona evaluadora y para poder reproducir la metodología en estudios futuros con tipos de texto o motores diferentes— también se ha diseñado una guía para la evaluación de estas pruebas. Véase el Anexo 2. Pautas para la evaluación «Precisión y fluidez».

Una vez hechas estas tres pruebas, tendremos la primera parte de la evaluación completada: la calidad. Así tendremos datos suficientes sobre cuál de los tres motores es mejor que el otro. Gracias a la primera prueba, sabremos cuál es el mejor motor, cuál es el segundo mejor y cuál es el peor motor de los escogidos para el presente estudio. Asimismo, gracias a la segunda y la tercera prueba, también sabremos cómo de fluidos o precisos son y tendremos más información para analizar si hay algún motor que gana

por goleada o, por el contrario, hay algún motor que es mejor en fluidez, pero que es peor en precisión o viceversa.

4.2.4. Productividad

En este punto, hay que proceder con la segunda parte de la evaluación humana: la productividad. Tal y como habíamos comentado anteriormente en las hipótesis del trabajo, era posible que los motores de TAN ofrecieran una calidad, precisión y fluidez mayor, pero que fueran más difíciles de poseditar. Para hacer esto, las herramientas DQF de TAUS nos ofrecen una prueba adicional.

En esta cuarta prueba también se utilizará la plataforma DQF de TAUS, la misma que en las tres pruebas anteriores. Las personas evaluadoras verán la oración en la LO, así como la traducción de un motor de TA, y tendrán que poseditar la traducción automática para conseguir una traducción de calidad humana. Esta prueba nos proporcionará los siguientes datos:

- Tiempo de edición: el tiempo que las personas evaluadoras tardan en hacer la posedición, así como el número promedio de palabras que se pueden hacer por hora.
- Distancia de edición: un número que dependerá de las modificaciones, ya sean adiciones o eliminaciones que la persona poseditora ha hecho durante la posedición. Cuantas más modificaciones haya hecho a la traducción en bruto propuesta por el motor de TA, mayor será el número. Si no ha hecho ninguna modificación, el número será 0.

Para llevar a cabo esta tarea, también se entregará una guía de posedición en la que se indicarán las instrucciones y los criterios a seguir en la misma. Asimismo, los motores se analizarán en dos grupos de estudio, ya que los objetivos son distintos:

1. Por una parte, se comparará el Traductor de Softcatalà con el Traductor de Google (ambos motores de TAN y, según las hipótesis del trabajo, ofrecerán una traducción de calidad mayor que Apertium; de esta manera, obtendremos

resultados sobre qué motor es el más útil para introducirse en un flujo de trabajo cuyo fin sea la diseminación en las combinaciones inglés-catalán).

2. Por otra parte, se comparará el Traductor de Softcatalà con Apertium, en colaboración con miembros de Softcatalà (ambos motores de código libre; de esta manera, sabremos qué motor deberían utilizar los miembros de Softcatalà para la traducción y localización de los proyectos en los que colaboren; este estudio también permitirá saber cuál es el motor de código libre en la combinación inglés-catalán puntero y que ofrece más calidad, a la vez que respeta la confidencialidad de los datos).

Como esta prueba es más costosa y los resultados serán más complejos de analizar, las personas evaluadoras se reducirán a 6 por grupo de estudio.

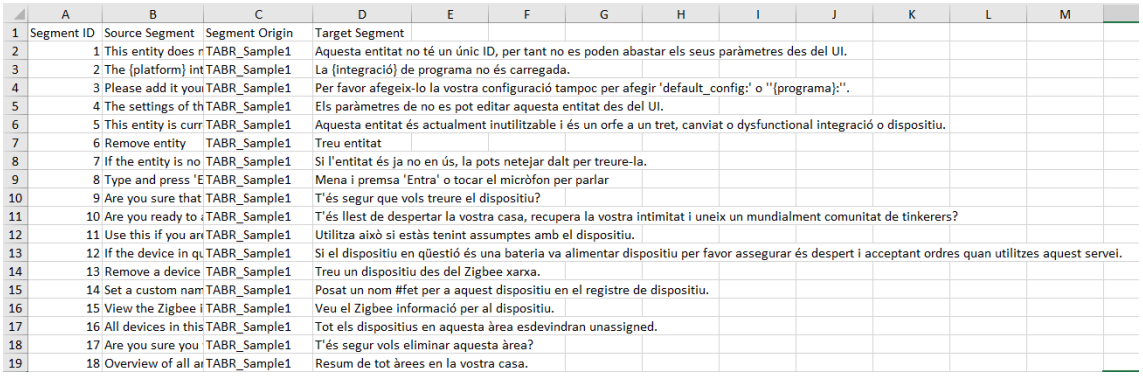
Sin embargo, esta prueba tiene una complicación más. Al tener que comparar la posedición de dos motores diferentes, no sería válido traducir el mismo texto, ya que si la persona poseditora 1 (PP1) trabaja con un único texto con dos motores distintos, la segunda posedición se vería beneficiada, ya que la persona encargada de la posedición conoce el texto (efecto de aprendizaje). En consecuencia, se ha decidido crear dos textos (T1 y T2) para evaluar en esta prueba de posedición. Dicho esto, esta prueba se ha organizado de la siguiente manera (para los dos grupos de estudio):

	T1 Motor 1	T2 Motor 1	T1 Motor 2	T2 Motor 2
PP1	x			x
PP2		x	x	
PP3	x			x
PP4		x	x	
PP5	x			x
PP6		x	x	

Tabla 1. Prueba de productividad: reparto de textos y motores

De esta manera, todas las personas evaluadoras utilizarán los dos motores que se estudiarán con textos diferentes y no habrá ninguna inclinación o preferencia en la posedición.

Para estas tres últimas pruebas, los archivos que acepta la plataforma también son XLS, XLSX, CSV o TMX, pero las columnas obligatorias cambian (ID del segmento, segmento en la LO, nombre del archivo en la LO y solo una propuesta de TA, ya que se evaluará individualmente). Véase la imagen a continuación:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Segment ID	Source Segment	Segment Origin	Target Segment									
2	1	This entity does r	TABR_Sample1	Aquesta entitat no té un únic ID, per tant no es poden abastar els seus paràmetres des del UI.									
3	2	The {platform} int	TABR_Sample1	La {integració} de programa no és carregada.									
4	3	Please add it you	TABR_Sample1	Per favor afegeix-lo la vostra configuració tampoc per afegir 'default_config:' o '{programa}:''.									
5	4	The settings of th	TABR_Sample1	Els paràmetres de no es pot editar aquesta entitat des del UI.									
6	5	This entity is curr	TABR_Sample1	Aquesta entitat és actualment inutilitzable i és un orfe a un tret, canviat o dysfunctional integració o dispositiu.									
7	6	Remove entity	TABR_Sample1	Treu entitat									
8	7	If the entity is no	TABR_Sample1	Si l'entitat és ja no en ús, la pots netejar dalt per treure-la.									
9	8	Type and press 'E	TABR_Sample1	Mena i premsa 'Entra' o tocar el micròfon per parlar									
10	9	Are you sure that	TABR_Sample1	T'és segur que vols treure el dispositiu?									
11	10	Are you ready to ¿	TABR_Sample1	T'és llest de despertar la vostra casa, recupera la vostra intimitat i uneix un mundialment comunitat de tinkerers?									
12	11	Use this if you ar	TABR_Sample1	Utilitza això si estàs tenint assumptes amb el dispositiu.									
13	12	If the device in qu	TABR_Sample1	Si el dispositiu en qüestió és una bateria va alimentar dispositiu per favor assegurar és despert i acceptant ordres quan utilitzes aquest servei.									
14	13	Remove a device	TABR_Sample1	Treu un dispositiu des del Zigbee xarxa.									
15	14	Set a custom nam	TABR_Sample1	Posat un nom #fet per a aquest dispositiu en el registre de dispositiu.									
16	15	View the Zigbee i	TABR_Sample1	Veu el Zigbee informació per al dispositiu.									
17	16	All devices in this	TABR_Sample1	Tot els dispositius en aquesta àrea esdevindran unassigned.									
18	17	Are you sure you	TABR_Sample1	T'és segur vols eliminar aquesta àrea?									
19	18	Overview of all ar	TABR_Sample1	Resum de tot àrees en la vostra casa.									

Figura 3. Fluency, Accuracy y Productivity: Ejemplo de un archivo preparado para la plataforma DQF de TAUS

5. Análisis de los resultados

En este apartado se expondrán los resultados obtenidos tras las evaluaciones humanas, de prueba en prueba, y se analizarán correspondientemente.

5.1. MT Ranking

Una vez todas las personas evaluadoras realizan la evaluación de MT Ranking, la plataforma DQF de TAUS permite descargar los resultados en archivos CSV o XLSX para poder analizarlos en profundidad y ver qué ha evaluado cada persona. No obstante, como nos interesa saber el resultado general, se han volcado todos los datos en una única hoja de Excel para poder trabajar con toda la información en un único lugar (véase la hoja de datos «MT Ranking – Resultados totales» en el Anexo 4. Hojas de datos del estudio). Esta valoración se ha llevado a cabo sobre el texto 1 (T1) de cada motor y, en

consecuencia, sobre 100 segmentos. Asimismo, la propia plataforma DQF ya ofrece algunas gráficas interesantes como la siguiente.

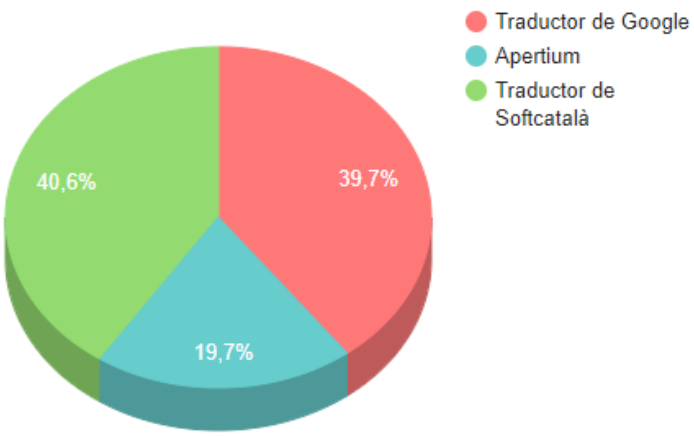


Figura 4. Resultados generales MT Ranking: porcentaje de veces que un motor ha recibido la puntuación 1

La imagen refleja qué motor consideran las personas evaluadoras que ofrece la mejor propuesta de traducción, a saber, qué propuesta necesitaría menos ediciones para obtener calidad humana. ¿Qué significan los porcentajes de la gráfica? Estos porcentajes indican las veces que los evaluadores han dado la puntuación 1 (la mejor propuesta de TA) a determinado motor. Por tanto, el Traductor de Softcatalà ha sido valorado como el mejor motor (un 40,6 % de las veces) frente al Traductor de Google (39,7 %) y Apertium (19,7 %). Pese a que la distancia entre el primer y el segundo motor es mínima —0,9 puntos porcentuales—, si se mira en profundidad la hoja de cálculos con todos los datos, se puede ver que, de las 11 personas evaluadoras, 10 han valorado que el Traductor de Softcatalà es el mejor y, la que no lo ha hecho, ha valorado que ambos motores están empatados (con una puntuación del 39 % como motores número 1 tanto para el Traductor de Softcatalà como el Traductor de Google).

Si observamos los datos de otra manera y creamos una tabla según el motor y, dentro de cada motor, indicamos el porcentaje de veces que se le ha dado cierta valoración —ya sea 1, 2 o 3—, obtenemos la siguiente gráfica.

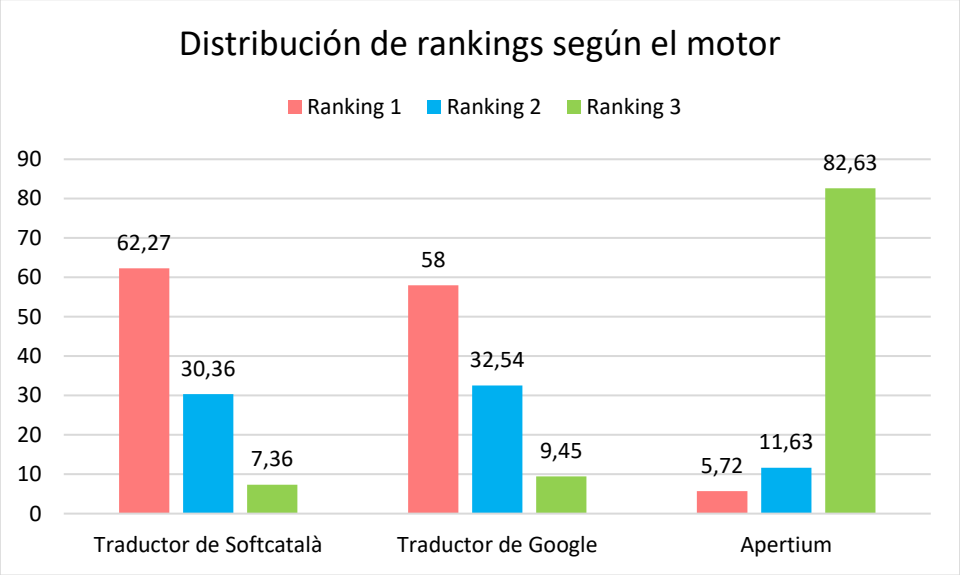


Figura 5. Resultados por motor MT Ranking: distribución de rankings

Estos datos y esta gráfica nos permiten ver que el Traductor de Softcatalà no solo ha obtenido más valoraciones como mejor motor (ranking 1), sino que también es el motor con menos valoraciones como peor motor (ranking 3). Aunque este tipo de evaluación y clasificación es relativa y nos permite ver qué motor es mejor pero no en qué grado (véase el apartado 3.4.2. La evaluación humana de la calidad), sí que se puede afirmar que, según la percepción de las personas evaluadoras y de manera casi unánime, el Traductor de Softcatalà es el que ofrece las mejores traducciones.

5.2. Fluidez y precisión

En segundo lugar y para acabar con el análisis de la calidad de traducción de los motores escogidos, se hará una revisión de los resultados de las pruebas de fluidez y precisión. Esta valoración se ha llevado a cabo sobre el texto 1 y el texto 2 de cada motor y, en consecuencia, se han evaluado 200 segmentos de cada motor. En esta prueba, las propuestas de traducción también eran anónimas y la persona evaluadora no conocía qué motor proporcionaba cada propuesta de traducción. Tras volcar todos los datos de la plataforma en única hoja de datos (véase la hoja de datos «Fluidez y precisión – Resultados totales» en el Anexo 4), se pueden obtener las siguientes gráficas.

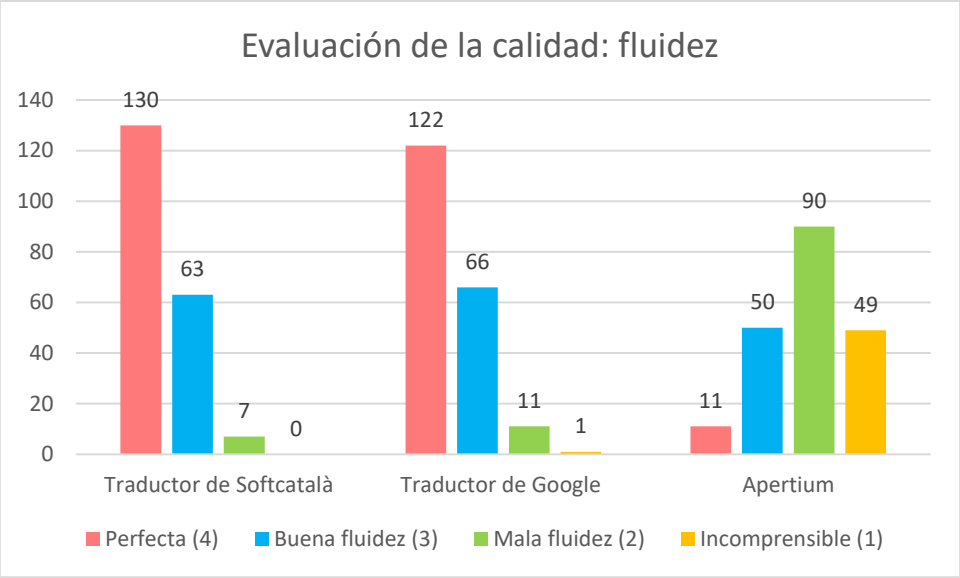


Figura 6. Evaluación de la calidad: fluidez

En esta primera gráfica, podemos observar la valoración de los 200 segmentos de cada motor, de conformidad con las guías y pautas de evaluación para la prueba de fluidez.

Como se puede observar, el Traductor de Softcatalà obtiene la mejor puntuación, ya que 130 segmentos tenían una fluidez perfecta y 63 segmentos una buena fluidez. Esto indica que 193 segmentos, un 96,5 % del total, ofrecen una fluidez perfecta o buena, permitiendo que el texto se lea de forma natural y de conformidad con las reglas de la LM —en este caso, el catalán—. El Traductor de Google tiene unos resultados muy similares, aunque un poco inferiores, con 122 segmentos valorados con una fluidez perfecta y 66 segmentos con una buena fluidez, obteniendo una fluidez perfecta o buena en un 94 % de los casos. En cambio, Apertium obtiene unos resultados muy diferentes, ya que solo tiene 11 y 50 segmentos valorados en las dos mejores categorías y únicamente un 30,5 % de los segmentos obtienen la valoración de fluidez perfecta o buena.

Pese a que la clasificación de MT Ranking es relativa, esta evaluación de fluidez nos permite ver que el Traductor de Softcatalà y el Traductor de Google ofrecen una calidad más similar que Apertium, ya que este último queda bastante más alejado en cuanto a fluidez. Mientras que el Traductor de Softcatalà y el Traductor de Google obtienen 0 y 1 segmentos con la puntuación 1 (Incomprensible), Apertium recibe esta puntuación en

49 segmentos. Esto también refuerza la hipótesis inicial del trabajo, ya que suponíamos que los motores de TAN ofrecen una fluidez perfecta o buena, y esta afirmación va en la línea de otros estudios del campo sobre TAN (véase el apartado 3.2. La traducción automática (TA).

En segundo lugar, y tras utilizar los resultados de precisión, podemos obtener la siguiente gráfica.

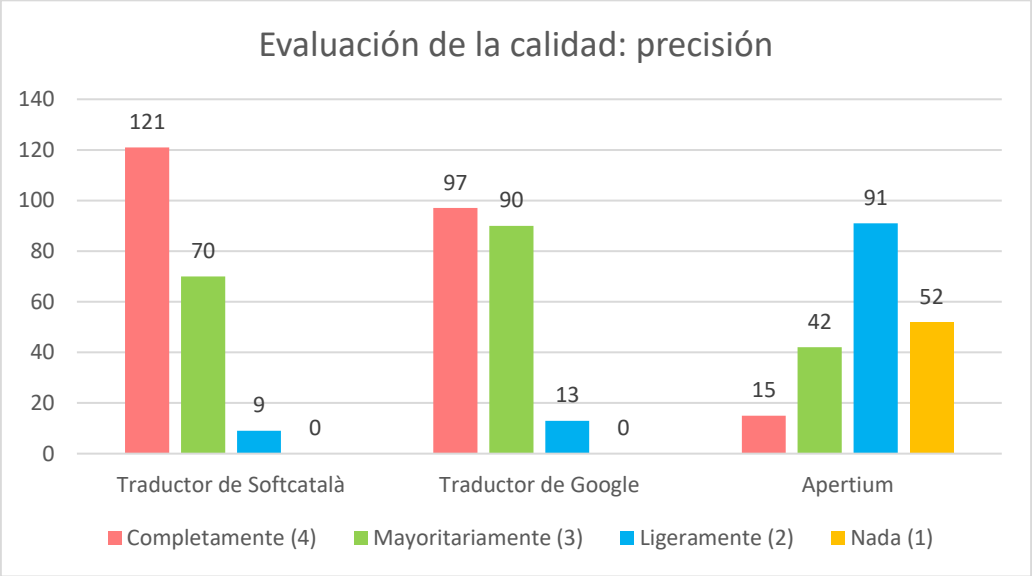


Figura 7. Evaluación de la calidad: precisión

Los datos de la prueba de precisión reflejan una realidad similar a la de fluidez: hay una gran diferencia de calidad entre el Traductor de Softcatalà y el Traductor de Google, por una parte, y Apertium, por la otra parte. Asimismo, en precisión, el Traductor de Softcatalà escala algunas posiciones y aumenta la distancia respecto al motor de Google a diferencia de los resultados de fluidez.

Más concretamente, el Traductor de Softcatalà y el Traductor de Google tienen 191 y 187 segmentos de 200 en las dos categorías de mayor calidad, respectivamente, las cuales equivalen a que la traducción en bruto propuesta contiene el significado del texto original completamente o mayoritariamente. No obstante, si se evalúa más en detalle, el Traductor de Softcatalà tiene 121 con la valoración 4 (Completamente), mientras que el Traductor de Google solo tiene 97; lo mismo pasa en la valoración 3 (Mayoritariamente), con 70 segmentos frente a 90. Esto indica que, pese a que ambos

motores de TA ofrezcan traducciones muy fluidas, casi perfectas en la mayoría de las situaciones y en un grado similar, el Traductor de Softcatalà es más preciso y acierta más que el Traductor de Google, alejándose de este en los resultados de precisión.

En cuanto a Apertium, 116 segmentos, un 58 % del total, han recibido la valoración 2 y 1 (Ligeramente y Nada, respectivamente), lo cual indica que las traducciones en bruto propuestas por el motor no recogen el significado del texto original y la precisión queda muy lejos de los otros dos motores del presente estudio.

5.3. Productividad

Una vez se ha hecho la evaluación de la calidad de los motores y sabemos qué motor ofrece una calidad de traducción mayor, se hará una revisión de los resultados de las posediciones, para así poder saber con cuál se es más productivo. Como esta evaluación se ha hecho con dos grupos de estudio diferentes, los pares de datos se analizarán independientemente.

Cuando las personas evaluadoras han hecho las posediciones, obtenemos el tiempo de posedición de cada segmento en milisegundos, así como la distancia de edición (cuantas más adiciones, eliminaciones o cambios a la traducción en bruto propuesta por el motor de TA, mayor es el número), para los dos textos de los pares de motores analizados. Para tratar estos datos, primero se han volcado los datos según el grupo de estudio (véase las hojas de datos «Evaluación de la productividad - Resultados globales [FOCUS GROUP 1]» y «Evaluación de la productividad - Resultados globales [FOCUS GROUP 2]» en el Anexo 4. Hojas de datos del estudio). A continuación, se han sumado los tiempos de posedición de todos los participantes para cada motor y se ha calculado la media (en milisegundos y segundos), así como la desviación estándar, para así analizar de manera estadística el tiempo de posedición medio por persona. Posteriormente, se ha hecho la prueba de Levene con Excel para ver si hay homocedasticidad, es decir, que las varianzas entre las muestras de los motores son iguales, ya que del cumplimiento de esta condición dependerá la formulación que empleemos en el contraste de medias. Si el p-valor resultante de la prueba de Levene es inferior a un cierto nivel de significación

(0,05), no hay homocedasticidad. A continuación, se utiliza la fórmula T de Student. Si el p-valor de esta segunda prueba es inferior a 0,05, podemos afirmar que las muestras son estadísticamente significativas. Asimismo, se ha procedido de la misma forma con la distancia de edición. De esta manera, podemos analizar los pares de datos de manera relevante e independiente.

Para profundizar el estudio, no solo se ha hecho un análisis del tiempo de posedición y de la distancia de edición general, sino que posteriormente también se han separado los segmentos en tres grupos según la longitud del texto en la LO: i) de 1 a 5 palabras (44 segmentos); ii) de 6 a 15 palabras (113 segmentos); y de 16 o >16 palabras (43 segmentos). Una vez hecha esta separación, los datos se han analizado estadísticamente de la misma manera que en el análisis general: cálculo de la media en milisegundos y segundos y desviación estándar, prueba de Levene y T de Student. Pese a que los datos originales están en milisegundos, en adelante se mostrarán en segundos para facilitar la comprensión.

5.3.1. Grupo de estudio 1 (Traductor de Softcatalà-Traductor de Google)

Para el primer grupo de estudio se han analizado las posediciones de las traducciones en bruto del Traductor de Softcatalà y el Traductor de Google. En este caso, se han volcado todos los datos de la plataforma DQF de TAUS en una única hoja de datos, tal como se menciona en el apartado anterior, y se obtienen los siguientes resultados generales.

	Traductor de Softcatalà		Traductor de Google	
	Media	Desviación	Media	Desviación
Tiempo de PE (s)	3909,077	21,616	4131,640	18,897
Distancia de edición (segmento)	9,797	13,872	10,358	13,926

Prueba de Levene (tiempo de PE)	0,665 (> 0,05)
T de Student (tiempo de PE)	0,197 (> 0,05)
Prueba de Levene (distancia de edición)	0,071 (> 0,05)
T de Student (distancia de edición)	0,42211267 (> 0,05)

Tabla 2. Resultados generales de productividad: grupo de estudio 1

Según la tabla anterior, podemos ver que el tiempo de posedición general es menor en el Traductor de Softcatalà que en el Traductor de Google (una diferencia de 222,563 segundos; un 5,69 % menor). En cuanto a la distancia de edición general, también podemos observar que la distancia de edición en el nivel del segmento es menor en el Traductor de Softcatalà (9,797) que en el Traductor de Google (10,358), con una diferencia total de 0,561 puntos. Además, la prueba de Levene da un resultado mayor a 0,05 en ambos casos (tiempo de posedición y distancia de edición) y, por tanto, hay homocedasticidad. No obstante, la T de Student arroja también valores superiores al 0,05, con lo que podemos afirmar que los resultados no son estadísticamente significativos. Por lo tanto, aunque se puede afirmar que, en términos generales y globales, el tiempo de posedición y la distancia de edición son menores en el Traductor de Softcatalà que en el Traductor de Google, no se puede afirmar que la diferencia sea estadísticamente significativa.

Para analizar estos datos más en profundidad, analizaremos estos resultados según la longitud de los segmentos (en palabras), tal como hemos mencionado anteriormente.

		1-5 palabras		6-15 palabras		16 o >16 palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Tiempo de PE (s)	Softcatalà	8,152	9,377	18,448	17,921	34,089	29,970
	Google	9,412	8,842	20,080	17,604	33,673	21,695

Prueba de Levene	0,966 (> 0,05)	0,799 (> 0,05)	0,046 (< 0,05)
T de Student	0,221893 (> 0,05)	0,124540331 (> 0,05)	0,877289313 (> 0,05)

Tabla 3. Resultados concretos de productividad: tiempo de posedición, grupo de estudio 1

Como se ve en la Tabla 3, el tiempo de posedición del Traductor de Softcatalà es menor en los segmentos de 1 a 5 palabras (8,152 segundos frente a 9,412 segundos) y de 6 a 15 palabras (18,448 segundos frente a 20,080 segundos). Estos resultados son la media de tiempo de posedición según la longitud del segmento. Además, la prueba de Levene da un resultado superior a 0,05 y, por lo tanto, hay homocedasticidad. La T de Student arroja para estos casos valores superiores a 0,05 y tampoco podemos afirmar que las diferencias sean estadísticamente significativas.

No obstante, en los segmentos largos de 16 o más de 16 palabras, el tiempo de posedición es inferior con el Traductor de Google (34,089 segundos frente a 33,673 segundos). La prueba de Levene da un resultado inferior a 0,05 y rechaza la hipótesis nula de varianzas iguales, por lo que no hay homocedasticidad. Asimismo, la T de Student también resulta superior a 0,05 y las diferencias tampoco son estadísticamente significativas en este caso.

		1-5 palabras		6-15 palabras		16 o >16 palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Distancia de edición (segmento)	Softcatalà	5,341	13,684	11,531	14,814	9,798	10,051
	Google	12,22	19,924	9,313	11,967	11,202	10,766

Prueba de Levene	0,001 (< 0,05)	0,006 (< 0,05)	0,255 (> 0,05)
T de Student	0,000835 (< 0,05)	0,007674 (< 0,05)	0,1977349 (> 0,05)

Tabla 4. Resultados concretos de productividad: distancia de edición, grupo de estudio 1

De conformidad con los resultados de la Tabla 4, la distancia de edición en el nivel del segmento es inferior en los segmentos de 1 a 5 palabras en el Traductor de Softcatalà (5,341 frente a 12,22). La prueba de Levene indica un resultado inferior a 0,05 y, por tanto, no hay homocedasticidad. No obstante, la T de Student arroja un p-valor inferior a 0,05 y en consecuencia podemos afirmar que la distancia de edición en los segmentos cortos es estadísticamente significativa.

En cuanto a los segmentos de longitud media, de 6 a 15 palabras, la distancia de edición es menor con el Traductor de Google (9,313 frente a 11,531), indicando la prueba de Levene y la T de Student (ambas por debajo de 0,05) que la diferencia también es estadísticamente significativa.

En cambio, para los segmentos de 16 o más de 16 palabras, la distancia de edición vuelve a ser inferior en el Traductor de Softcatalà (9,798 frente a 11,202). La prueba de Levene arroja un resultado de 0,255 (superior a 0,05) y, por lo tanto, no hay homocedasticidad. A continuación, la T de Student también indica un p-valor superior a 0,05 y, en consecuencia, tampoco podemos afirmar que esta diferencia en la distancia de edición sea estadísticamente significativa.

Para resumir las tablas anteriores y simplificar el análisis de este primer grupo de estudio, en la comparación del tiempo de posedición y la distancia de edición entre el Traductor de Softcatalà y el Traductor de Google, véase la tabla a continuación con un

resumen global de los resultados. El símbolo del asterisco «*» indica que el p-valor es inferior a 0,05, es decir, que los resultados son estadísticamente significativos.

	Tiempo de PE	Distancia de edición
1-5 palabras	Softcatalà < Google	Softcatalà < Google*
5-15 palabras	Softcatalà < Google	Softcatalà > Google*
16 o >16 palabras	Softcatalà > Google	Softcatalà < Google

Tabla 5. Resumen de resultados de productividad: tiempo de PE y distancia de edición, grupo de estudio 1

De conformidad con la Tabla 5 y como resumen de los resultados del grupo de estudio 1, podemos afirmar que el tiempo de posedición es menor con el Traductor de Softcatalà con respecto al Traductor de Google en los segmentos de longitud corta y media (de 1 a 5 palabras y de 5 a 15 palabras, respectivamente), aunque es mayor en los segmentos de longitud larga (16 o más de 16 palabras). No obstante, estos resultados no son estadísticamente significativos y sería recomendable aumentar las muestras y volver a efectuar las pruebas.

En cambio, en cuanto a la distancia de edición, podemos afirmar que esta es muy superior para el Traductor de Google en los segmentos de 1 a 5 palabras, pero que es un poco inferior en los segmentos de 5 a 15 palabras. Estos resultados son estadísticamente significativos. No obstante, en los segmentos de 16 o más de 16 palabras, la distancia de edición es menor para el Traductor de Softcatalà, aunque sería recomendable aumentar las muestras y volver a hacer las pruebas para obtener resultados estadísticamente significativos.

5.3.2. Grupo de estudio 2 (Traductor de Softcatalà-Apertium)

Para este segundo grupo de estudio, se han analizado las posediciones de las traducciones en bruto del Traductor de Softcatalà y Apertium. Para hacerlo, se ha

repetido el mismo procedimiento que el utilizado en el análisis del grupo de estudio 1. Se han volcado los datos de la plataforma DQF de TAUS en una única hoja de datos, tal como se menciona en el apartado anterior, y se han hecho las pruebas estadísticas correspondientes, con el objetivo de obtener las tablas a continuación.

	Traductor de Softcatalà		Apertium	
	Media	Desviación	Media	Desviación
Tiempo de PE (s)	1859,517	12,572	3743,413	21,846
Distancia de edición (segmento)	6,810	12,682	24,850	20,201

Prueba de Levene (tiempo de PE)	3,49E-21 (< 0,05)
T de Student (tiempo de PE)	2,42E-39 (< 0,05)
Prueba de Levene (distancia de edición)	1,006E-26 (< 0,05)
T de Student (distancia de edición)	1,171E-135 (< 0,05)

Tabla 6. Resultados generales de productividad: grupo de estudio 2

De conformidad con la Tabla 6, podemos ver que el tiempo de posedición general es menor en el Traductor de Softcatalà que en Apertium (una diferencia de 1883,897 segundos; un 101,31 % menor). En cuanto a la distancia de edición general, también podemos observar que la distancia de edición en el nivel del segmento es menor en el Traductor de Softcatalà (6,810) que en Apertium (24,850). En este último caso, la diferencia es considerable: 18,04 puntos. Además, la prueba de Levene da un resultado menor a 0,05 en ambos casos (tiempo de posedición y distancia de edición). En adición, la T de Student arroja también valores inferiores al 0,05, con lo que podemos afirmar que los resultados son estadísticamente significativos, aunque a simple vista se ve que hay grandes diferencias, no como en el grupo de estudio anterior. Por lo tanto, se puede afirmar que, en términos generales y globales, el tiempo de posedición y la distancia de edición son mucho menores en el Traductor de Softcatalà que en Apertium. Esta afirmación puede hacerse de manera estadísticamente significativa.

Para analizar estos datos más en profundidad y, al igual que con el primer grupo de estudio, analizaremos estos resultados según la longitud de los segmentos (en palabras).

		1-5 palabras		6-15 palabras		16 o >16 palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Tiempo de PE (s)	Softcatalà	5,956	6,149	14,187	15,804	25,836	16,218
	Apertium	14,111	14,884	28,643	18,552	55,700	28,095

Prueba de Levene	0,001 (< 0,05)	0,00002 (< 0,05)	9,577E-07 (< 0,05)
T de Student	4,49E-08 (< 0,05)	2,434E-26 (< 0,05)	3,8408E-19 (< 0,05)

Tabla 7. Resultados concretos de productividad: tiempo de posedición, grupo de estudio 2

Como puede observarse en la Tabla 7, el tiempo de posedición del Traductor de Softcatalà es muy inferior en todos los grupos de segmentos: tanto en los de 1 a 5 palabras (5,956 segundos frente a 14,111 segundos; un 136,91% inferior) como en los de 6 a 15 palabras (14,187 segundos frente a 28,643 segundos; un 101,89% inferior) y los de 16 o más de 16 palabras (25,836 segundos frente a 55,700 segundos; un 115,58% inferior). Además, la prueba de Levene da un resultado inferior a 0,05 y, por lo tanto, no hay homocedasticidad. La T de Student arroja para estos casos valores inferiores a 0,05 y podemos afirmar que las diferencias sean estadísticamente significativas.

		1-5 palabras		6-15 palabras		16 o >16 palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Distancia de edición (segmento)	Softcatalà	6,212	11,854	10,732	13,400	10,116	8,702
	Apertium	40,659	29,534	37,761	18,137	36,155	13,398

Prueba de Levene	4,86E-26 (< 0,05)	5,105E-08 (< 0,05)	0,0001 (< 0,05)
T de Student	9,709E-26 (< 0,05)	9,535E-83 (< 0,05)	3,937E-44 (<0,05)

Tabla 8. Resultados concretos de productividad: distancia de edición, grupo de estudio 2

De conformidad con los resultados de la Tabla 8, la distancia de edición en el nivel del segmento es inferior en los segmentos de 1 a 5 palabras en el Traductor de Softcatalà (6,212 frente a 40,659). La prueba de Levene indica un resultado inferior a 0,05 y, por tanto, no hay homocedasticidad. Asimismo, la T de Student arroja un p-valor inferior a 0,05 y en consecuencia podemos afirmar que la distancia de edición en los segmentos cortos es estadísticamente significativa.

En cuanto a los segmentos de longitud media —de 6 a 15 palabras— y de longitud larga —de 16 o más de 16 palabras—, la distancia de edición continúa siendo menor con el Traductor de Softcatalà (10,732 frente a 37,761 y 10,116 frente a 36,155, respectivamente), indicando la prueba de Levene y la T de Student (ambas por debajo de 0,05) que la diferencia también es estadísticamente significativa. Aunque la diferencia en estos dos grupos de segmentos es menor, continúa siendo extremadamente amplia.

Al igual que en el primer grupo de estudio, se ha hecho una tabla adicional como resumen de los datos anteriores y así simplificar el análisis de este segundo grupo de estudio, tanto en la comparación del tiempo de posedición como de la distancia de edición entre el Traductor de Softcatalà y Apertium. Véase la tabla a continuación con un resumen global de los resultados. El símbolo del asterisco «*» indica que el p-valor es inferior a 0,05, es decir, que los resultados son estadísticamente significativos.

	Tiempo de PE	Distancia de edición
1-5 palabras	Softcatalà < Apertium*	Softcatalà < Apertium*
5-15 palabras	Softcatalà < Apertium*	Softcatalà > Apertium*
16 o >16 palabras	Softcatalà > Apertium*	Softcatalà < Apertium*

Tabla 9. Resumen de resultados de productividad: tiempo de PE y distancia de edición, grupo de estudio 2

A diferencia del primer grupo de estudio, donde había un motor que ganaba ligeramente por encima del otro en la mayoría de los aspectos (pese a que algunos resultados no eran estadísticamente significativos), en este segundo grupo de estudio queda muy claro que hay una gran diferencia entre los motores. Los resultados indican que el Traductor de Softcatalà tiene tiempos de posedición y distancias de edición mucho inferiores que Apertium en todos los grupos de segmentos (tanto en los cortos como en los de longitud media y los de longitud larga). Además, estas diferencias son estadísticamente significativas.

6. Conclusiones

Así pues, tras haber analizado todos los resultados obtenidos en las diferentes pruebas y evaluaciones que se han llevado a cabo en el presente estudio, en este apartado se tratarán las conclusiones y justificaciones de los resultados, de conformidad con los objetivos iniciales planteados, así como las posibles líneas de investigación que han surgido de este trabajo.

En cuanto al primer objetivo (***1. Revisar la bibliografía más actualizada sobre motores de TA y su evaluación***), el presente trabajo define claramente la situación actual de la TA y su evaluación. La evaluación automática de los motores de TA es importante para los desarrolladores de motores, pero no es determinante para declarar qué motor ofrece una calidad de traducción o una productividad mayor. En consecuencia, en el presente trabajo se han llevado a cabo una serie de evaluaciones humanas para solventar este problema y cumplir con los otros objetivos del estudio.

En cuanto al segundo objetivo (***2. Analizar qué método y sistema de TA escogido [Traductor de Softcatalà, Traductor de Google, Apertium] ofrece una calidad de traducción mayor***), se han llevado a cabo las pruebas MT Ranking, Fluidez y Precisión de DQF de TAUS.

En lo referente a la prueba MT Ranking, se ha decidido de forma casi unánime que el Traductor de Softcatalà es el mejor motor de los tres evaluados, ya que, según la

percepción de 10 de las 11 personas, este motor ha conseguido la mejor puntuación en la mayoría de los casos. La persona que no ha valorado que el Traductor de Softcatalà es mejor que el resto indica que el Traductor de Softcatalà y el Traductor de Google están empatados.

Respecto a la prueba de fluidez, el Traductor de Softcatalà y Google obtienen resultados similares, ya que un 96,5 % y un 94 % de sus segmentos reciben una valoración de fluidez perfecta o buena. En cambio, Apertium queda muy lejos de estos resultados y solo recibe esta valoración en un 30,5 % de sus segmentos. El resto, un 69,5 % del total, recibe las valoraciones de fluidez mala o incomprensible. Por lo que corresponde a la prueba de precisión, el Traductor de Softcatalà continúa en primera posición y se distancia ligeramente del Traductor de Google, aunque ambos obtienen muy buenos resultados. Apertium, una vez más, queda muy rezagado en resultados, obteniendo valoraciones pésimas en comparación con los otros dos motores.

Tras estas tres pruebas, se puede afirmar que el Traductor de Softcatalà es el motor de TA que ofrece una calidad de traducción mayor (tanto en percepción de las personas evaluadoras, como en fluidez y precisión), quedando el Traductor de Google en segundo lugar y Apertium en tercera posición.

En lo que respecta al tercer objetivo (***3. Descubrir qué motor de TA escogido ofrece un rendimiento mayor al poseditar, con el objetivo de saber cuál puede aportar más beneficios al introducirlo en un flujo de traducción profesional o aficionado***), se ha llevado a cabo un análisis de la productividad a través de la posedición que dos grupos de estudio han hecho de dos textos diferentes.

Por una parte, al analizar los resultados del primer grupo de estudio (Traductor de Softcatalà-Traductor de Google), se puede afirmar que, en términos globales y generales, el tiempo de posedición y la distancia de edición es menor para el Traductor de Softcatalà que para el Traductor de Google. Es decir, se traduce más texto en menos tiempo con el Traductor de Softcatalà y, además, hay que hacer menos modificaciones (adiciones, ediciones u omisiones) a la traducción en bruto propuesta por dicho motor.

Ahora bien, es cierto que, en algunos casos, los resultados no son estadísticamente significativos y sería recomendable aumentar el tamaño de las pruebas (ya sea aumentando el número de segmentos o la cantidad de personas evaluadoras). Por otra parte, al analizar los resultados del segundo grupo de estudio (Traductor de Softcatalà-Apertium), se puede afirmar rotundamente que el tiempo de posedición y la distancia de edición es mucho menor para el Traductor de Softcatalà que para Apertium. Además, estos resultados son estadísticamente significativos en todos los supuestos analizados. En adición, se cumple la hipótesis de que los motores de TAN son más precisos y fluidos que los motores de TABR y, aunque tal vez los errores estén más escondidos y sean más difíciles de encontrar, la posedición de motores de TAN continúa siendo más productiva que la de motores de TABR.

En consecuencia, y tras analizar los datos globalmente, se puede asegurar que el Traductor de Softcatalà ofrece un rendimiento por encima del resto de motores evaluados. De esta manera, este motor aportará más beneficios a la hora de introducir TA en un flujo de trabajo cuyo objetivo sea la diseminación, ya sea de manera profesional o aficionada. Así pues, también puede afirmarse que los motores de TA «in-domain» o entrenados con corpus específicos, como el Traductor de Softcatalà, ofrecen una calidad o un rendimiento mayor a los motores de TA genéricos o «out-of-domain», como el Traductor de Google.

Sobre el cuarto y último objetivo (***4. Sentar las bases para una investigación futura y un posible doctorado***), los resultados abren una serie de preguntas que pueden ser interesantes de abordar en estudios futuros:

- ¿Reafirmaría y reforzaría su superioridad el Traductor de Softcatalà frente al Traductor de Google al aumentar la cantidad de segmentos y de personas evaluadoras?
- ¿Es el Traductor de Softcatalà el mejor motor de TA en otros ámbitos? Ya que solo se ha probado en localización de software («in-domain»), puede ser

interesante evaluar sectores como el jurídico o el médico (donde sería «out-of-domain»).

- Los datos obtenidos pueden utilizarse para analizar otros aspectos muy interesantes. Además de profundizar el estudio al separar los segmentos según la longitud del texto en la LO, ¿qué resultados obtendríamos si también se separan los segmentos según su calidad, de conformidad con los resultados de fluidez y precisión? ¿Qué cambios hacen las personas poseditoras en los segmentos de mayor calidad? ¿Y en los de menor calidad? Por lo tanto, podría analizarse el comportamiento de las personas poseditoras según la calidad del segmento.
- Además, para mejorar la calidad del Traductor de Softcatalà, ¿es más útil aumentar la cantidad de textos para entrenar el motor o mejorar la calidad de los que ya hay? ¿Qué hay que priorizar, el volumen de textos o la calidad de los textos a la hora de entrenar el motor?
- Asimismo, habría sido interesante hacer una comparación y un estudio de las evaluaciones de productividad con el Traductor de Softcatalà entre las personas poseditoras del grupo de estudio 1 y del grupo de estudio 2. De esta manera, podríamos ver también datos no solo entre motores, sino también entre sujetos (cómo se comportan, qué modifican y qué aceptan traductores profesionales y aficionados/voluntarios). No obstante, ha sido imposible hacerlo por limitaciones temporales, pero se tiene en cuenta para estudios futuros.

Aunque ahora hay nuevos caminos abiertos, el presente trabajo ha dejado claro que el Traductor de Softcatalà es actualmente el mejor motor de TA disponible para la combinación inglés-catalán en localización de software. Además, si se tiene en cuenta que es de código libre, no solo tiene beneficios de calidad y productividad, sino también de precio, limitación de caracteres, traducción de documentos grandes y confidencialidad, entre otros elementos.

7. Bibliografía

Ahrenberg, L. (2017). Comparing Machine Translation and Human Translation: A Case Study. *Proceedings of the Workshop on Human-Informed Translation and Interpreting Technology*, 21-28. https://doi.org/10.26615/978-954-452-042-7_003

ALPAC. (s. f.). *Language and machines*. 138.

Arthern, P. J. (1979). *MACHINE TRANSLATION AND COMPUTERIZED TERMINOLOGY SYSTEMS A TRANSLATOR'S VIEWPOINT.* 32.

Bahdanau, D., Cho, K., y Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*. http://arxiv.org/abs/1409.0473

Banerjee, S., y Lavie, A. (2005). *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. 8.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., y Vilar, J.-M. (2009). Statistical Approaches to Computer-Assisted Translation. *Computational Linguistics*, *35*(1), 3-28. https://doi.org/10.1162/coli.2008.07-055-R2-06-29

Bender, O., Hasan, S., Vilar, D., Zens, R., y Ney, H. (2005). Comparison of Generation Strategies for Interactive Machine Translation. *Conference Proceedings*, 8.

Bentivogli, L., Bisazza, A., Cettolo, M., y Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: A Case Study. *arXiv:1608.04631 [cs]*. http://arxiv.org/abs/1608.04631

Bojar, OndAŽej, Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., y Monz, C. (2018). *Findings of the 2018 Conference on Machine Translation (WMT18)*. 37.

Bojar, Ondřej, Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., … Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 131-198. https://doi.org/10.18653/v1/W16-2301

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., y Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16*(2), 79–85.

Cadwell, P., Castilho, S., O'Brien, S., y Mitchell, L. (2016). Human factors in machine translation and post-editing among institutional translators. *Translation Spaces*, *5*(2), 222-243. https://doi.org/10.1075/ts.5.2.04cad

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., y Schroeder, J. (2007). (Meta-) evaluation of machine translation. *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, 136-158. https://doi.org/10.3115/1626355.1626373

Castaño, M. A., Casacuberta, F., y Vidal, E. (1997). *Machine translation using neural networks and finite-state models*. 8.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., y Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of*

Mathematical Linguistics, *108*(1), 109-120. https://doi.org/10.1515/pralin-2017-0013

Chan, S.-W. (2014). *Routledge Encyclopedia of Translation Technology* (1.ª ed.). Routledge. https://doi.org/10.4324/9781315749129

Crego, J., y Senellart, J. (2016). Neural Machine Translation from Simplified Translations. *arXiv:1612.06139 [cs]*. http://arxiv.org/abs/1612.06139

Cronin, M. (2012). *Translation in the Digital Age* (1.ª ed.). Routledge. https://doi.org/10.4324/9780203073599

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research* -, 138. https://doi.org/10.3115/1289189.1289273

Doğru, G., Martín-Mor, A., y Ortiz-Rojas, S. (2019). *MTradumàtica: Free Statistical Machine Translation Customisation for Translators*. 3.

Etchegoyhen, T., Fernández Torné, A., Azpeitia, A., Martínez Garcia, E., y Matamala, A. (2018, mayo). Evaluating Domain Adaptation for Machine Translation Across Scenarios. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1002

Fields, P., Hague, D. R., Koby, G. S., Lommel, A., y Melby, A. (2014). What Is Quality? A Management Discipline and the Translation Industry Get Acquainted. *Tradumàtica: Tecnologies de La Traducció*, *12*, 404. https://doi.org/10.5565/rev/tradumatica.75

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

Forcada, M. L., y Ñeco, R. P. (1997). *Recursive Hetero-associative Memories for Translation*.

Ginestí, M., y Forcada, M. L. (2009). *LA TRADUCCIÓ AUTOMÀTICA EN LA PRÀCTICA: APLICACIONS, DIFICULTATS I ESTRATÈGIES DE DESENVOLUPAMENT*. 18.

Görög, A. (2014). Quantifying and benchmarking quality: The TAUS Dynamic Quality Framework. . . *ISSN*, 12.

Graham, Y., Baldwin, T., Moffat, A., y Zobel, J. (2013). *Continuous Measurement Scales in Human Evaluation of Machine Translation*. 9.

Green, S. (2016). *Interactive Machine Translation*. 93.

Guerberof-Arenas, A. (2008). *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. *1*, 11.

Han, L. (2018). Machine Translation Evaluation Resources and Methods: A Survey. *arXiv:1605.04515 [cs]*. http://arxiv.org/abs/1605.04515

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., … Zhou, M. (2018). Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv:1803.05567 [cs]*. http://arxiv.org/abs/1803.05567

He, Y., y Way, A. (2009). *Learning Labelled Dependencies in Machine Translation Evaluation*. 9.

Hearne, M., y Way, A. (2011). Statistical Machine Translation: A Guide for Linguists and Translators: SMT for Linguists and Translators. *Language and Linguistics Compass*, *5*(5), 205-226. https://doi.org/10.1111/j.1749-818X.2011.00274.x

Hutchins, W. J. (1986). *Machine Translation: Past, Present, Future*.

ISO (International Organization for Standardization). (2015a). *ISO 9001:2015 Quality management systems—Requirements*.

ISO (International Organization for Standardization). (2015b). *ISO 17100:2015 Translation Services—Requirements for Translation Services*.

ISO (International Organization for Standardization). (2017). *ISO 18587:2017 Translation Services—Post-editing of Machine Translation Output—Requirements*.

Jean, S., Cho, K., Memisevic, R., y Bengio, Y. (2015). On Using Very Large Target Vocabulary for Neural Machine Translation. *arXiv:1412.2007 [cs]*. http://arxiv.org/abs/1412.2007

Junczys-Dowmunt, M., y Grundkiewicz, R. (2016). Phrase-based Machine Translation is State-of-the-Art for Automatic Grammatical Error Correction. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1546-1556. https://doi.org/10.18653/v1/D16-1161

Koby, G. S., Fields, P., Hague, D. R., Lommel, A., y Melby, A. (2014). Defining Translation Quality. *Tradumàtica: Tecnologies de La Traducció*, *12*, 413. https://doi.org/10.5565/rev/tradumatica.76

Koehn, P. (2017). Neural Machine Translation. *ArXiv:1709.07809 [Cs]*. http://arxiv.org/abs/1709.07809

Koehn, P., y Haddow, B. (2009). *Interactive assistance to human translators using statistical machine translation methods*. 8.

Koehn, P., y Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. *Proceedings of the Workshop on Statistical Machine Translation - StatMT '06*, 102. https://doi.org/10.3115/1654650.1654666

Koehn, P., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., y Moran, C. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07*, 177. https://doi.org/10.3115/1557769.1557821

Langlais, P., Foster, G., y Lapalme, G. (2000). TransType: A Computer-Aided Translation Typing System. *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*. https://www.aclweb.org/anthology/W00-0507

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., y Toral, A. (2020). A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. *Journal of Artificial Intelligence Research*, *67*, 653–672-653–672. https://doi.org/10.1613/jair.1.11371

Läubli, S., Sennrich, R., y Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4791–4796. https://doi.org/10.18653/v1/D18-1512

Leiva Rojo, J. (2018). Aspects of human translation: The current situation and an emerging trend. *Hermēneus. Revista de Traducción e Interpretación*, *20*, 257-294. https://doi.org/10.24197/her.20.2018.257-294

López-Pereira, A. (2019). *Traducción automática neuronal y traducción automática estadística: Percepción y productividad*. 19.

Martín-Mor, A. (2017). *MTradumàtica: Statistical Machine Translation Customisation for Translators*. 15.

Martín-Mor, A., Sánchez-Gijón, P., y Piqué, R. (2016). *Tradumàtica: Tecnologies de la traducció*.

Melby, A., Fields, P., Hague, D. R., Koby, G. S., y Lommel, A. (2014). Defining the Landscape of Translation. *Tradumàtica: Tecnologies de La Traducció*, *12*, 392. https://doi.org/10.5565/rev/tradumatica.74

Melby, A. K. (1978). Design and implementation of a computer-assisted translation system. *Équivalences*, *9*(2), 37-48. https://doi.org/10.3406/equiv.1978.1014

Moorkens, J., Castilho, S., Gaspari, F., y Doherty, S. (Eds.). (2018). *Translation Quality Assessment: From Principles to Practice* (Vol. 1). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7

Moorkens, J., y O'Brien, S. (2017). *Assessing User Interface Needs of Post-Editors of Machine Translation*. 22.

Nitzke, J. (2019). *Problem solving activities in post-editing and translation from scratch*. Zenodo. https://doi.org/10.5281/ZENODO.2546446

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*, *25*(3), 197-215. https://doi.org/10.1007/s10590-011-9096-7

O'Brien, S. (2012). *Towards a dynamic quality evaluation model for translation*.

O'Brien, S., y Moorkens, J. (2014). *Towards Intelligent Post-Editing Interfaces*. 9.

Papineni, K., Roukos, S., Ward, T., y Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on*

Association for Computational Linguistics - ACL '02, 311. https://doi.org/10.3115/1073083.1073135

Peris, Á., y Casacuberta, F. (2018). Active Learning for Interactive Neural Machine Translation of Data Streams. *arXiv:1807.11243 [cs]*. http://arxiv.org/abs/1807.11243

Peris, Á., y Casacuberta, F. (2019). Online Learning for Effort Reduction in Interactive Neural Machine Translation. *arXiv:1802.03594 [cs]*. http://arxiv.org/abs/1802.03594

Popel, M. (2018). CUNI Transformer Neural MT System for WMT18. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 482–487. https://doi.org/10.18653/v1/W18-6424

Sánchez-Gijón, P. (2014). La investigación en traducción y calidad, cosa de dos. . . *ISSN*, 6.

Sánchez-Gijón, P., Moorkens, J., y Way, A. (2019). Post-editing neural machine translation versus translation memory segments. *Machine Translation*, *33*(1-2), 31-59. https://doi.org/10.1007/s10590-019-09232-x

Schwab, K. (s. f.-a). *The Fourth Industrial Revolution*. 172.

Schwab, K. (s. f.-b). *The Global Competitiveness Report 2019*. 666.

Shannon, C., y Weaver, W. (s. f.). *The Mathematical Theory of Communication*. 131.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., y Makhoul, J. (2016). *A Study of Translation Edit Rate with Targeted Human Annotation*. 9.

Snover, M. G., Madnani, N., Dorr, B., y Schwartz, R. (2009). TER-Plus: Paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, *23*(2-3), 117-127. https://doi.org/10.1007/s10590-009-9062-9

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., y Sawaf, H. (1997). *ACCELERATED DP BASED SEARCH FOR STATISTICAL TRANSLATION*. 5.

Tomás, J., y Casacuberta, F. (2006). Statistical phrase-based models for interactive computer-assisted translation. *Proceedings of the COLING/ACL on Main Conference Poster Sessions -*, 835-841. https://doi.org/10.3115/1273073.1273180

Way, A. (2018). Quality Expectations of Machine Translation. En J. Moorkens, S. Castilho, F. Gaspari, y S. Doherty (Eds.), *Translation Quality Assessment* (Vol. 1, pp. 159-178). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_8

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv:1609.08144 [Cs]*. http://arxiv.org/abs/1609.08144

Ye-Yi Wang, Acero, A., y Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 577-582. https://doi.org/10.1109/ASRU.2003.1318504

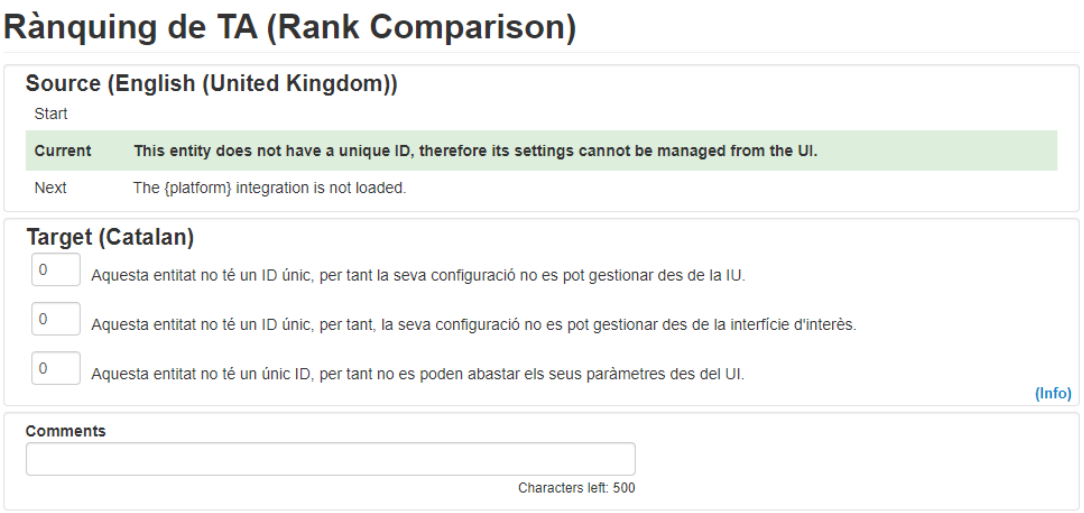
Anexo 1. Pautas para la evaluación «MT Ranking»

Objectiu

L'objectiu d'aquest document és definir clarament les pautes i instruccions per fer la tasca d'avaluació humana de la traducció automàtica. Aquesta tasca d'avaluació consisteix a fer una valoració relativa de les propostes de traducció de diferents motors de TA.

Descripció de la prova

1. Rebreu un correu electrònic amb l'enllaç al sistema d'avaluació i unes instruccions breus. En fer clic a l'enllaç, anireu a la pàgina on es fa l'avaluació.
2. En aquesta pàgina, feu clic al botó "Start" per començar.
3. La tasca consisteix en l'avaluació de 100 segments. A la interfície d'usuari veureu "Source" amb l'oració en anglès i "Target" amb tres oracions en català. Cadascuna d'aquestes oracions prové d'un motor de traducció automàtica anònim. Vegeu la imatge a continuació.



4. Heu de llegir totes les traduccions i fer una valoració relativa d'aquests motors. Què significa això? Heu d'assignar les següents puntuacions:

1	És la millor traducció proposada de les tres que s'ofereixen
2	És la segona millor traducció proposada de les tres que s'ofereixen
3	És la pitjor traducció proposada de les tres que s'ofereixen

5. En cas d'empat de dues oracions, podeu puntuar les dues primeres oracions amb el número 1 i deixar la tercera oració amb el 3, o viceversa.
6. Si voleu, podeu utilitzar el camp "Comments" per fer aclariments (opcional).
7. Podeu tornar a segments anteriors fent clic al botó "Previous".
8. Podeu interrompre l'avaluació fent clic a "Home".
9. Si no heu tancat la sessió del navegador, feu clic a "Continue" per reprendre l'avaluació.
10. Si heu tancat la sessió del navegador, feu clic a l'enllaç del correu electrònic automàtic i, a continuació, feu clic a "Continue" per reprendre l'avaluació.

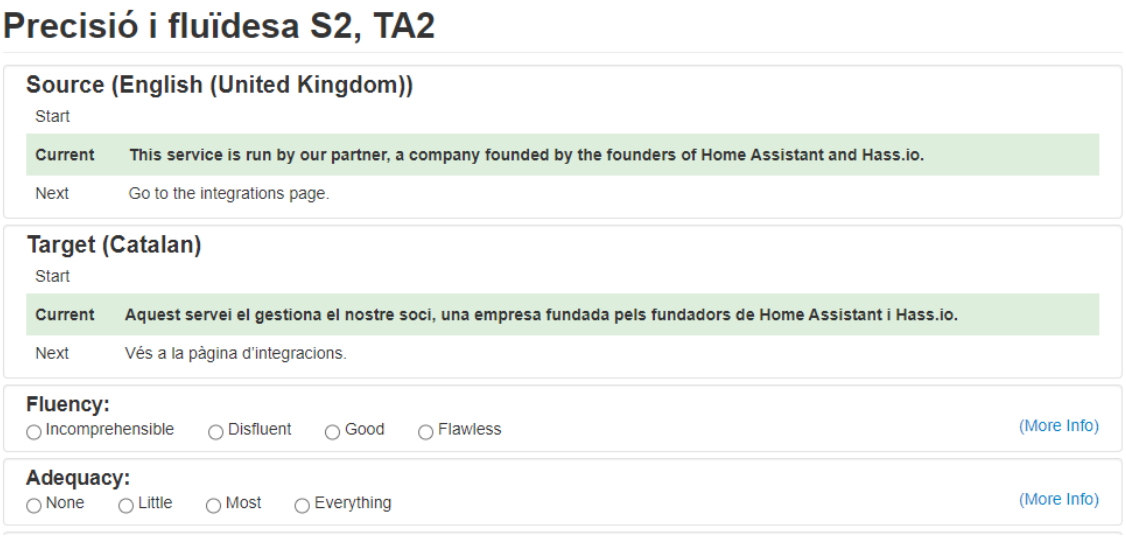
Anexo 2. Pautas para la evaluación «Precisión y fluidez»

Objectiu

L'objectiu d'aquest document és definir clarament les pautes i instruccions per fer la tasca d'avaluació humana de la traducció automàtica. Aquesta tasca d'avaluació consisteix a fer una valoració de la precisió i de la fluïdesa que ofereixen diferents traduccions en brut de diversos motors de TA.

Descripció de la prova

1. Rebreu un correu electrònic amb l'enllaç al sistema d'avaluació i unes instruccions breus. En fer clic a l'enllaç, anireu a la pàgina on es fa l'avaluació.
2. En aquesta pàgina, feu clic al botó "Start" per començar.
3. La tasca consisteix en l'avaluació de 100 segments. A la interfície d'usuari veureu "Source" amb l'oració en anglès i "Target" amb una oració en català. Aquesta oració prové d'un motor de traducció automàtica anònim. Vegeu la imatge a continuació.



4. Heu de llegir la proposta de traducció de motor de TA i, a continuació, hi heu de fer una valoració de la fluïdesa. Què significa això? Heu d'assignar les següents puntuacions:

Incomprehensible	L'oració està escrita molt malament i no es pot entendre res.
Disfluent	L'oració està escrita molt malament i costa molt entendre-la.
Good	L'oració s'entén parcialment, ja que és fluïda, però té diversos errors petits.
Flawless	L'oració és molt fluïda i no té errors.

5. Heu de llegir la proposta de traducció de motor de TA i, a continuació, hi heu de fer una valoració de precisió. Què significa això? Heu d'assignar les següents puntuacions:

None	El significat de la traducció proposada no conte res de què parla el text en la llengua origen.
Little	El significat de la traducció proposada conté fragments que sí que respecten el text en la llengua origen.
Most	El significat de la traducció proposada respecta quasi en la seva totalitat el text en la llengua origen.
Everything	El significat de la traducció proposada respecta totalment el text en la llengua origen. No hi ha ni addicions ni omissions.

Anexo 3. Pautas para la evaluación «Productividad»

Objectiu

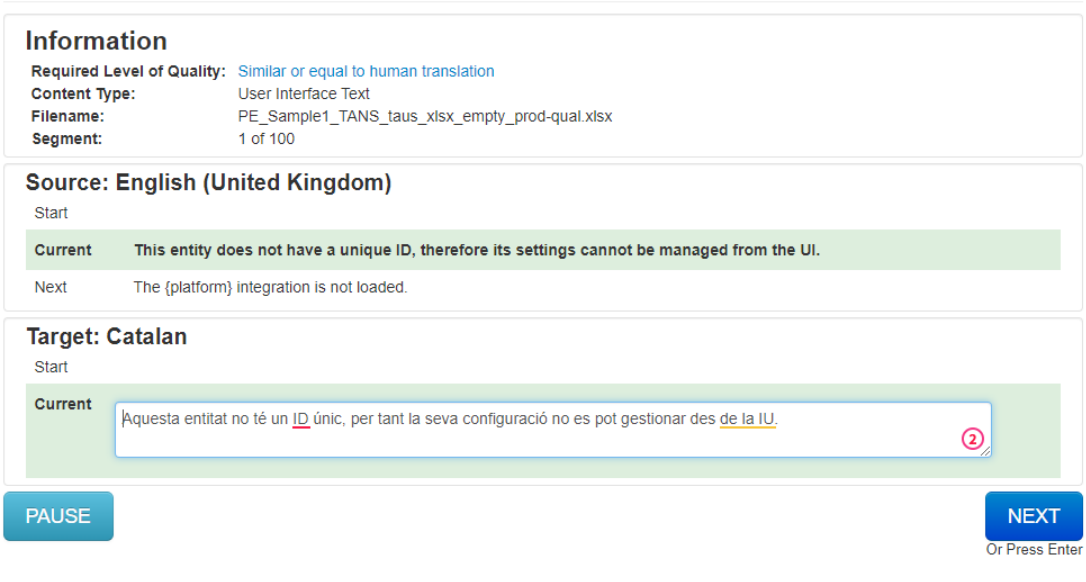
L'objectiu d'aquest document és definir clarament les expectatives de qualitat i les pautes de postedició per a l'avaluació de motors de traducció automàtica (TA).

IMPORTANT!

En aquesta tasca, haureu de posteditar una sèrie de segments amb TA. L'objectiu és calcular quant de temps tardareu en fer la postedició (per tant, se us cronometrarà).
Per tant, és recomanable que feu la postedició com si es tractés d'un encàrrec professional, tot d'una, ja que es comptaran els milisegons i el número d'edicions per poder comparar l'esforç d'edició d'un motor A i un motor B.

Podeu posar pausa a la tasca, però no és recomanable, per tal d'evitar errors i tergiversar el càlcul de temps de la postedició. A més, quan passeu de segment, no podreu tornar enrere, així que el que hàgiu fet es desarà i no podrà canviar-se.

La interfície de postedició serà la que veieu a la captura de pantalla següent, on podreu veure el segment actual en el text origen (anglès) i el segment en text meta (català).



Expectatives de qualitat

Traducció automàtica (TA)

La TA és un procés on el text en la llengua origen es tradueix a la llengua meta automàticament. El text que surt d'un motor de TA s'anomena "traducció en brut". En molts casos, la traducció en brut es postedita, en diferents mesures, per millorar la seva qualitat.

Postedició

La postedició (PE) és el procés d'editar i corregir la traducció en brut d'un motor de TA per millorar la seva qualitat i assegurar que el text meta reprodueix fidelment el text original amb una qualitat humana.

Pautes generals

- No s'ha d'esborrar el text en brut completament.
- S'ha d'aprofitar al màxim la traducció en brut.
- S'ha de comprovar que la traducció en brut correspon i reprodueix fidelment l'original.
- No s'ha d'ometre o afegir informació important.
- S'ha de comprovar si els termes que no s'han traduït han de quedar-se sense traduir.
- S'han de tenir en compte les normes gramaticals i sintàctiques de la llengua meta.
- Tingueu en compte (si és possible) els errors que més es repeteixen per comentar-los posteriorment. D'aquesta manera, es pot intentar millorar el motor de TA.

Pautes pràctiques

Tractament (vós/tu):

Quan és l'usuari qui s'adreça a l'ordinador, s'utilitzarà sempre l'imperatiu en segona persona del singular («tu»). Això ho trobarem sovint en els menús, en alguns quadres de diàleg i en els botons.

Original	Edit
Incorrecte	Editar
Correcte	Edita

Quan és l'ordinador qui s'adreça a l'usuari per a donar-li informació, fer-li una pregunta, etc., la forma que cal utilitzar és l'imperatiu en segona persona del plural («vós»). Això ho trobem sobretot en la documentació i en alguns quadres de diàleg.

Original	Choose the language you want to show this website in.
Incorrecte	Tria la llengua en què vols mostrar aquest lloc web.
Correcte	Trieu la llengua en què voleu mostrar aquest lloc web.

Convenció geogràfica i local (números, mesures, divises, dates, etc.):

Problemes de localització que poder estar causats perquè el motor de TA no ha sabut adaptar la traducció a la convenció geogràfica corresponent. Documenteu-vos per seguir les normes de la llengua meta.

Original	More than 250,000.75 people died because of this pandemic.
Incorrecte	Més de 250,000.75 persones van morir a causa d'aquesta pandèmia.
Correcte	Més de 250.000,75 persones van morir a causa d'aquesta pandèmia.

Problemes de localització:

El traductor automàtic no ha respectat algun dels elements claus en la localització, com poden ser les variables, els enllaços o elements no traduïbles dins d'una cadena que s'ha de traduir.

Original	Hello {name}, welcome.
Incorrecte	Hola, {nom}, benvingut.
Correcte	Hola, {name}, et donem la benvinguda.

Problemes de precisió i traduccions incorrectes:

El text en la llengua meta no representa de manera acurada el text original. Les paraules o termes que s'han fet servir són incorrectes o massa literals, o no s'ha respectat un nom propi.

Original	Home Assistant information
Incorrecte	Informació de l'assistent d'estar per casa
Correcte	Informació de Home Assistant

Errors de puntuació

La TA sovint tendeix a imitar la puntuació del text original. S'han de comprovar les normes de puntuació de la llengua meta amb cura.

Original	Nobody appears to have asked this "behavioural insights team."
Incorrecte	Ningú sembla haver preguntat a aquest "equip de coneixement de comportament."
Correcte	Sembla que ningú ha preguntat a aquest «equip de coneixement del comportament».

Ús de majúscules

L'anglès tendeix a utilitzar més majúscules que el català. Per tant, s'ha de seguir l'ús natural de les majúscules en català, independentment del que digui la TA.

Original	When Prime Minister Boris Johnson held his first major press conference.
Incorrecte	El Primer Ministre Boris Johnson va celebrar la seva primera gran conferència de premsa.
Correcte	El primer ministre Boris Johnson va celebrar la seva primera gran conferència de premsa.

Anexo 4. Hojas de datos del estudio

Enlace para consultar «HomeAssistant – Completo»: https://bit.ly/2UkvHad.

Enlace para consultar «MT Ranking – Resultados totales»: https://bit.ly/2XDDpyn.

Enlace para consultar «Fluidez y precisión – Resultados totales»: https://bit.ly/2YdXnPa.

Enlace para consultar «Evaluación de la productividad - Resultados globales (FOCUS GROUP 1)»: https://bit.ly/3fNRdg0.

Enlace para consultar «Evaluación de la productividad - Resultados globales (FOCUS GROUP 2)»: https://bit.ly/37NPB32.