Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the article:

López Pereira, Ariana; Sánchez-Gijón, Pilar, dir. Traducción automática neuronal y traducción automática estadística : percepción y productividad. 2018. 83 p.

This version is available at https://ddd.uab.cat/record/202425

under the terms of the BY-SA license

Traducción automática neuronal y traducción automática estadística: percepción y productividad

Trabajo de Fin de Máster

Autora: Ariana López Pereira
Tutora: María Pilar Sánchez Gijón
Máster en Tradumática: Tecnologías de la Traducción

Facultad de Traducción e Interpretación
Universitat Autònoma de Barcelona
Curso 2017-2018

Datos del Trabajo de Fin de Máster

Título: Traducción automática neuronal y traducción automática estadística: percepción y productividad

Autora: Ariana López Pereira

Tutora: María Pilar Sánchez Gijón

Centro: Facultad de Traducción e Interpretación

Estudios: Máster en Tradumática: Tecnologías de la Traducción

Curso académico: 2017-2018

Resumen: gracias al gran progreso que ha experimentado el campo de la traducción automática (TA) en los últimos años, es necesario revisar el uso y la percepción de esta por parte de los traductores. El objetivo principal de este trabajo es determinar la percepción y la productividad, en términos de tiempo y número de ediciones, de un grupo de traductores al utilizar sistemas de traducción automática estadística (TAE) y traducción automática neuronal (TAN). Este proyecto se centra en cómo diez traductores, todos ellos con experiencia profesional, perciben estos dos sistemas para conocer cuál prefieren. Asimismo, se busca obtener datos reales de los tiempos y distancias de posedición. Para conseguirlo, seis de los diez traductores realizaron varias con las herramientas del Dynamic Quality Framework (DQF) utilizando los sistemas de traducción automática de Google Neural Machine y Microsoft Translator (TAE) en dos textos diferentes de inglés a español, un manual de instrucciones y una página web de marketing. Los resultados mostraron que los traductores prefieren considerablemente el motor de TAN sobre el de TAE. Asimismo, los resultados prueban que la distancia de edición es inferior para los segmentos poseditados con TAN, pero el esfuerzo de edición, en tiempo, es mucho mayor que el de TAE.

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

Ariana López Pereira
Máster en Tradumática
Universitat Autònoma de Barcelona

Abstract: Thanks to the great progress seen in the machine translation (MT) field in recent years, the use and perception of MT by translators need to be revisited. The main objective of this paper is to determine the perception and productivity (in terms of time and number of editings) of a group of translators when using Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) systems. This presentation is focused on how ten professional translators perceive these two systems in order to know which one they prefer. The aim is also to obtain real data regarding the edit distance and post-editing time when working with these two systems. In order to do so, several tests were performed by six out of the ten translators with the Dynamic Quality Framework (DQF) tools (MT Ranking and productivity tasks) using Google Neural Machine Translation (NMT) and Microsoft Translator (SMT) APIs in two different English into Spanish texts, an instruction manual and a marketing webpage. Results showed that translators considerably prefer NMT over SMT. Likewise, results prove that the edit distance is lower for the segments post-edited with NMT, but the post-editing time, is much higher than in SMT.

Palabras clave (en): Neural Machine Translation, Statistical Machine Translation, productivity, edit distance, post-editing time, perception, post-editing.

Índice

1 Introducción 8
1.1 Objetivos e hipótesis 9
2 Marco teórico y antecedentes 11
2.1 Historia de la traducción automática 11
2.2 Traducción automática 12
2.2.1 Aplicaciones de la traducción automática 12
2.2.2 Tipos de motor de traducción automática 13
2.2.3 Traducción automática estadística 16
2.2.4 Traducción automática neuronal 16
2.3 Evaluación de la calidad de la traducción automática 17
2.3.1 Escalas de medida automáticas 18
2.3.2 Intervención humana en la evaluación de la traducción automática 20
2.3.3 Normas y modelos para la evaluación de la calidad 21
2.4 Posedición 24
2.4.1 Flujo de trabajo (TA + PE) 24
2.4.2 Tipos de posedición 26
2.4.3 Modos de integración de la TA en el flujo de traducción 27
2.4.4 La traducción automática en el flujo profesional 28
3 Marco metodológico 32
3.1 Descripción de las pruebas 33
3.2 Preparación de las pruebas 35
3.3 Instrumentos empleados 35
3.3.1 Instrumentos en común 35
3.3.2 Instrumentos para cada prueba 36
3.3.3 Sujetos 38

3.4 Prueba 1: MT Ranking 42
3.5 Prueba 2: prueba de productividad 44
3.5.1 Evaluación de calidad 46
4 Resultados obtenidos 49
4.1 Análisis de los resultados 49
4.2 Interpretación de los resultados 53
4.2.1 MT Ranking 53
4.2.2 Prueba de evaluación de calidad 54
4.2.3 Prueba de productividad 56
5 Conclusiones 73
5.1 MT Ranking 73
5.2 Pruebas de productividad 74
6 Bibliografía 76
7 Anexos 82

Índice de tablas

Tabla 1: desglose de participantes para las pruebas de productividad 45
Tabla 2: descripción de la escala de grados de fluidez de DQF 47
Tabla 3: descripción de la escala de grados de precisión de DQF..... 48
Tabla 4: descripción de resultados posibles para la categoría de mayor calidad 51
Tabla 5: descripción de resultados posibles para la categoría de calidad intermedia 52
Tabla 6: descripción de resultados posibles para la categoría de calidad baja ... 52
Tabla 7: resumen de los datos estadísticos obtenidos para la categoría de mayor calidad para la distancia de edición 58
Tabla 8: resumen de los datos estadísticos obtenidos para la categoría de mayor calidad para el tiempo de posedición..... 60
Tabla 9: resumen de los datos recogidos de la categoría de mayor calidad 62
Tabla 10: resumen de los datos estadísticos obtenidos para la categoría de calidad intermedia para la distancia de edición 63
Tabla 11: resumen de los datos estadísticos obtenidos para la categoría de calidad intermedia para el tiempo de posedición..... 65
Tabla 12: resumen de los datos recogidos de la categoría de calidad intermedia 67
Tabla 13: resumen de los datos estadísticos obtenidos para la categoría de calidad baja para la distancia de edición..... 68
Tabla 14: resumen de los datos estadísticos obtenidos para la categoría de calidad baja para el tiempo de posedición 70
Tabla 15 resumen de los datos recogidos de la categoría de calidad baja..... 72

Índice de figuras

Figura 1: imagen del triángulo de Vauquois 13
Figura 2: esquema de errores de DQF 22
Figura 3: plataforma en línea de Google Cloud 40
Figura 4: plataforma en línea de Microsoft Azure 40
Figura 5: ventana de confguración del complemento en Sdl Trados Studio 41
Figura 6: hoja de datos de la memoria de traducción para la prueba de MT Ranking 43
Figura 7: pantalla de inicio de una de las pruebas de calidad en DQF 47
Figura 8: aviso de TAUS de finalización de una prueba 48
Figura 9: gráfica de la prueba de MT Ranking 54
Figura 10: gráfica de puntuación de la prueba de evaluación de fluidez (en %) 55
Figura 11: gráfica de puntuación de la prueba de evaluación de precisión (en %) 55
Figura 12: distancia de edición de la categoría de mayor calidad, menos de cinco palabras 57
Figura 13: distancia de edición de la categoría de mayor calidad, de seis a diecinueve palabras 57
Figura 14: distancia de edición de la categoría de mayor calidad, más de veinte palabras 57
Figura 15: tiempo de posedición (en ms) de la categoría de mayor calidad, menos de cinco palabras 60
Figura 16: tiempo de posedición (en ms) de la categoría de mayor calidad, de seis a diecinueve palabras 60
Figura 17: tiempo de posedición (en ms) de la categoría de mayor calidad, más de veinte palabras 60
Figura 18: distancia de edición de la categoría de calidad intermedia, menos de cinco palabras 63
Figura 19: distancia de edición de la categoría de calidad intermedia, de seis a diecinueve palabras 63
Figura 20: distancia de edición de la categoría de calidad intermedia, más de veinte palabras 63

Figura 21: tiempo de posedición (en ms) de la categoría de calidad intermedia, menos de cinco palabras 65
Figura 22: tiempo de posedición (en ms) de la categoría de calidad intermedia, de seis a diecinueve palabras 65
Figura 23: tiempo de posedición (en ms) de la categoría de calidad intermedia, más de veinte palabras 65
Figura 24: distancia de edición de la categoría de calidad baja, menos de cinco palabras 68
Figura 25: distancia de edición de la categoría de calidad baja, de seis a diecinueve palabras 68
Figura 26: distancia de edición de la categoría de calidad baja, más de veinte palabras 68
Figura 27: tiempo de posedición (en ms) de la categoría tres, menos de cinco palabras 70
Figura 28: tiempo de posedición (en ms) de la categoría tres, de seis a diecinueve palabras 70
Figura 29: tiempo de posedición (en ms) de la categoría tres, más de veinte palabras 70

1 Introducción

En los últimos años, tanto el campo de la traducción como la industria que lo acompaña han experimentado grandes cambios. La necesidad que surge de traducir textos, de localizar productos o, sencillamente, de comunicarse de alguna manera, es cada vez mayor. Simultáneamente, existe la necesidad de que estas tareas se realicen de la forma más productiva posible, mientras se busca la forma de que los costes sean cada vez inferiores. Así, la traducción automática ha surgido como una solución a estos problemas. Esto se puede comprobar, por ejemplo, en el informe ProjecTA del año 2015. Asimismo, el empleo de la traducción automática ha generado nuevas formas de usarla. Entre estas cabe destacar el *light post-editing*, que implica el menor número de ediciones posibles para que el texto sea comprensible, o la traducción desatendida, que consiste en aplicar traducción automática en bruto, sin poseditar.

Los cambios que está viviendo el campo de la traducción y, es especial, de la traducción automática, hace que sea más necesario trabajar con herramientas que permiten una alta productividad sin poner en riesgo la calidad. Para ello, este trabajo tiene por objetivo realizar una aportación al ámbito profesional, basándose en resultados reales, y distinguir las diferencias que se presentan al trabajar con un motor de traducción automática neuronal y con uno de estadística. Asimismo, esta aportación pretende ampliar la investigación ya existente en este campo sobre las diferencias de estos dos motores. De esta forma, tras obtener los resultados de las distintas pruebas que se presentan en este trabajo, se podrá encauzar la investigación de una forma diferente.

En el ámbito personal, este proyecto tiene por objetivo ampliar mis competencias como investigadora, así como los conocimientos relacionados con la traducción automática, para emplearlos posteriormente en el mundo laboral. Asimismo, mediante este proyecto se pretende abrir una posible vía a realizar un Doctorado en Traducción y Nuevas Tecnologías.

Por último, es necesario mencionar que los resultados expuestos en este trabajo ya se presentaron en la conferencia anual de la European Association for Machine Translation, celebrada en mayo de 2018 en Alicante.

1.1 Objetivos e hipótesis

Como se mencionaba anteriormente, el presente trabajo tiene como objetivo explorar distintas características de uso de los motores de traducción automática neuronal y traducción automática estadística. En primer lugar, se pretende determinar la percepción de los traductores con el uso de los distintos motores de traducción automática. De esta forma, se busca saber si la percepción de que la traducción automática neuronal es superior y si esa tecnología de vanguardia supone una revolución tanto para el campo de la traducción como para el mercado (Torres Hostench et al, 2016). Para explorar las percepciones de los traductores, se usará una prueba de MT Ranking del DQF de TAUS, como se verá más adelante.

Durante los últimos años se han realizado pruebas acerca de la percepción de los traductores (Koponen, 2012) y se ha tratado el esfuerzo de posedición en la traducción automática estadística (Guerberof, 2011). No obstante, aquí se pretenden abordar ambos escenarios, tanto el de la percepción de los traductores a la hora de utilizar los distintos motores como si estas sensaciones son correctas. Se busca saber, entonces, si la percepción inicial de los traductores se corresponde con los resultados obtenidos a partir de sus pruebas.

En segundo lugar, este trabajo busca evaluar la productividad a la hora de usar un sistema de traducción automática neuronal y uno de traducción automática estadística. Para ello, se pretende descubrir si la productividad, en términos de tiempo y distancia de edición, es mayor al utilizar un motor u otro. Para ello, se comprobará cuánto tiempo se tarda en poseditar tanto un texto de marketing como un texto extraído de un manual de instrucciones con un motor de TAN como con uno de TAE. Tal y como recoge Koponen (2010), «post-editing time has been a commonly used measure of post-editing effort (Krings, 2001; O'Brien, 2005; Specia et al, 2009; Tatsumi, 2009; Tatsumi and Roturier, 2010; Specia, 2011; Carl et al, 2011)». La variable introducida de

los diferentes tipos de texto[1] busca responder si existen diferencias, tanto en tiempo como en distancia de edición, entre el uso de sistemas de TAN o de TAE en un texto de marketing y un manual de uso.

De forma paralela, se buscan obtener los resultados relacionados con la calidad de los sistemas de traducción automática. Se pretende resolver la hipótesis de que el sistema de traducción automática neuronal es más fluido que el de estadística, mientras que el sistema de traducción automática estadística es mucho más preciso. Para ello se utilizará una prueba de evaluación de calidad de DQF, que establece una escala de puntuación del 1 al 4, siendo 1 nada preciso o fluido y 4 totalmente fluido y preciso.

Para finalizar, tras realizar todas estas pruebas, se pretende extraer las conclusiones de si, por un lado, la percepción de los traductores son las mismas que proporcionan los resultados, esto es, si el motor que ellos consideran que proporciona mejores resultados e implica un menos esfuerzo de trabajo es con el que se obtienen mejores resultados (mayor productividad y menor esfuerzo de posedición).

[1] El motivo que ha llevado a escoger únicamente estos dos tipos de texto es que son de los tipos más comunes que se busca traducir ahora mismo en el mercado de la traducción. Lo ideal sería tratar cuantas más tipologías fuera posible para determinar si existe una adecuación de cada una de ellas para cada motor.

2 Marco teórico y antecedentes

A la hora de definir el marco teórico de este trabajo, se pueden diferenciar varias líneas. En primer lugar, se tratará la historia y los antecedentes de la traducción automática. Posteriormente, se trabajará con la traducción automática, los distintos tipos de motores y la evaluación de calidad de esta. Finalmente, nos centraremos en la posedición, las diferentes formas de posedición, así como el flujo de trabajo a la hora de emplear traducción automática y posedición.

2.1 Historia de la traducción automática

En los años sesenta, tras los avances tecnológicos experimentados a raíz de la segunda guerra mundial, y con el ideal de FAHQT todavía lejos, se creó el Automatic Language Processing Advisory Committee (ALPAC), un comité formado por ocho científicos cuyo propósito era evaluar las necesidades en traducción automática del gobierno de Estados Unidos y el estado de la cuestión en aquel momento, centrándose especialmente en la combinación de ruso a inglés. El comité llegó a la conclusión de que la oferta de traductores era mucho mayor que la demanda y aseveró que no había necesidad para traducir una gran cantidad de textos que se estaban traduciendo (Hutchins, 1992).

Con este panorama, se determinó que solo se podría seguir investigando si se llegaba a alcanzar un sistema de traducción de alta calidad sin la intervención humana (el ideal FAHQT). No obstante, a día de hoy, se está de acuerdo en que el estudio tenía fallos: se centró en la comparación de las traducciones en bruto de los tres sistemas de traducción automática con tres traducciones realizadas por traductores de un mismo texto. Los errores fueron la presentación de los textos, los criterios de evaluación de las traducciones y la elección de los traductores participantes. Se llegó a la conclusión de que era un gasto innecesario y, de esta forma, los investigadores no se centraron en las posibles ventajas y beneficios del uso de la traducción automática. Así, se abandonaron los estudios, reemplazándolos por estudios en lingüística teórica (Hutchins, 1986).

Tras la publicación del informe ALPAC, hubo un largo periodo en el que no se realizaron avances en el campo de la traducción automática. Sin embargo, en el campo de la traducción se continuó avanzando incesantemente. En lo que respecta a la traducción automática, todo mejoró con el procesamiento de palabras, que facilitaba el trabajo con los textos, ya que se podían utilizar corpus ya existentes y permitía la posedición directamente en pantalla. Así, se evitaban las correcciones de una segunda persona o teclear de nuevo el texto. Con el paso del tiempo, comenzaron a aparecer nuevos sistemas de traducción automática comerciales.

2.2 Traducción automática

Es esencial definir en primer lugar la traducción automática. Una de las definiciones más completas la da Ginestí y Forcada:

> «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». (Ginestí y Forcada, 2009)

Así, la traducción automática es el proceso de traducción que emplea un sistema informático (ya sean ordenadores o programas) y produce una traducción en bruto. La traducción automática se engloba dentro del campo de la lingüística computacional, que el diccionario Oxford define como «the branch of linguistics in which the techniques of computer science are applied to the analysis and synthesis of language and speech».

2.2.1 Aplicaciones de la traducción automática

Según sus aplicaciones, Ginestí y Forcada (2009) distinguen dos formas de uso de la traducción automática, a saber:

- Comprensión o asimilación: consiste en generar un texto traducido solo para obtener una idea general. En este caso, se suele desconocer la lengua de origen. Para esta aplicación no es necesario que el texto no tenga errores, ya que es suficiente si el texto final resulta comprensible. Algunos ejemplos de este uso es la traducción automática de sitios web (que permite una comprensión general de

lo que se está mostrando en la página en ese momento) o la traducción de documentación o correspondencia interna de una empresa.

- Publicación o diseminación: aquí la traducción automática es un paso intermedio en la producción de un texto meta que será publicado. El texto traducido mediante traducción automática tiene que revisarse, esto es, *poseditarse* (más adelante se verá esta definición). Prácticamente cualquier texto puede ser objeto de esta aplicación, pero es necesario tener en cuenta que, según la temática, se obtendrán mejores o peores resultados. Por ejemplo, en la actualidad, se ha demostrado que los textos altamente especializados poseditados con traducción automática neuronal producen errores (Luong et al, 2015).

2.2.2 Tipos de motor de traducción automática

Según su arquitectura lingüística, esto es, cómo esté construido el motor de traducción automática, se pueden distinguir los tipos motores que se verán a continuación. Así, el triángulo de Vauquois (Chan, 2015) representa los distintos motores según su arquitectura:

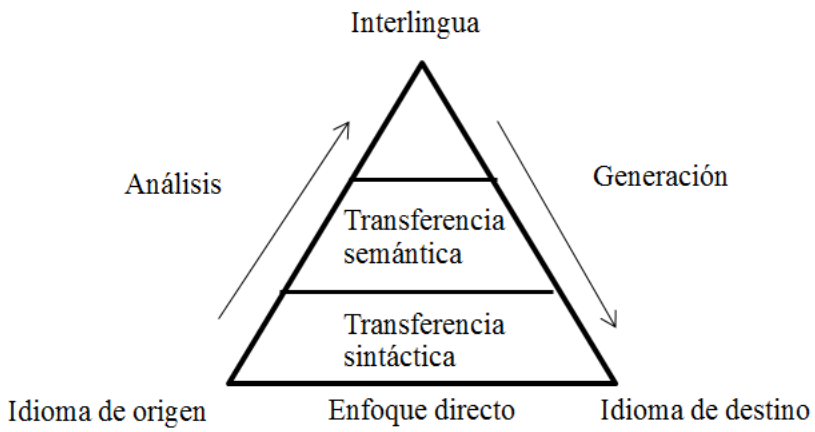


Figura 1: imagen del triángulo de Vauquois

Según su arquitectura computacional, se pueden distinguir los siguientes motores:

- Traducción automática basada en diccionarios (*dictionary based machine translation*): fue el método usado entre los años cuarenta y sesenta. Se basaba en diccionarios electrónicos bilingües; este sistema todavía es útil para traducir frases, pero no oraciones (Tripathi y Sarkhel, 2010).
- Traducción automática basada en reglas (TABR): estos sistemas utilizan diccionarios bilingües y se constituyen mediante la construcción de reglas lingüísticas. Aunque se adaptan muy fácilmente a las circunstancias, las excepciones en las reglas gramaticales suponen un importante problema. Según su arquitectura lingüística, pueden distinguirse las siguientes construcciones para un motor de TABR:
    - Enfoque directo o *direct approach*: el texto de origen se traduce directamente, sin pasar por ningún tipo de intermediario. En este tipo de enfoque, se eliminan las flexiones de género y número y se parte directamente de las formas base, para después buscarlas en un diccionario y traducirlas a partir de ahí (Chéragui, 2012).
    - *Transfer based*: en este caso, el texto de origen pasa por una fase intermedia en la que se vuelve un concepto abstracto. Tripathi y Sarkhel (2010) lo explican así: «source language is transformed into an abstract, less language-specific representation. An equivalent representation (with same level of abstraction) is then generated for the target language using bilingual dictionaries and grammar rules». Usa tres componentes principales: análisis, transferencia y síntesis. En el análisis, se procesa tanto la estructura sintáctica o la semántica, según corresponda; en la transferencia, se transfiere la estructura sintáctica/semántica del idioma meta; por último, en la síntesis, se reproducen los constituyentes del idioma de origen al idioma meta.
    - Interlingua: este enfoque es el más útil para los sistemas multilingües. Consta de dos fases, una de análisis y otra de generación:

> In Interlingua, source language is transformed into an auxiliary/intermediary language (representation) which is independent of any of the languages involved in the translation. The translated verse for the target language is

> then derived through this auxiliary representation. Hence, only two modules i.e., analysis and synthesis are required in this type of system. Also, because of its independency on the language pair for translation, this system has much relevance in multilingual machine translation. This emphasizes on single representation for different languages. (Tripathi y Sarkhel, 2010).

- Traducción automática basada en corpus (*Corpus based machine translation*, CBMT): desde 1989, es el sistema que más predomina. En este campo, destaca sobre todo la traducción automática estadística. El primero en mencionarla fue Warren Weaver, en 1949. A día de hoy, es uno de los sistemas más extendidos y consiste en usar grandes corpus bilingües paralelos para crear el sistema de traducción. Dentro de este tipo, se pueden distinguir varios modelos, según en qué esté basado el sistema:
    - Basado en la palabra
    - Basado en la frase
    - Basado en la sintaxis
    - Basado en ejemplos
    - Basado en el contexto
- Enfoque híbrido: en estos casos, se mezcla el enfoque de la traducción basada en corpus con la de transferencia: «The main idea in this approach is to automatically learn syntactic transfer rules from limited amounts of word-aligned data. This data contains all the needed information for parsing, transfer, and generation of the sentences» (Chéragui, 2010).

Asimismo, cuando se habla de motores de traducción automática, es importante mencionar que existe una variante a estos, conocida como «sublenguaje». Un sublenguaje se define como un «semi-autonomous, complex semiotic systems, based on and derived from language [whose] use presupposes special education and is restricted to communication along specialists in the same or closely related fields» (Sager et al 1980). Algunas ventajas del uso de un sublenguaje son que permite disponer de una terminología clara, se reduce la homografía y los problemas gramaticales que se presentan son los propios del campo que se está tratando (Hutchins, 1992).

2.2.3 Traducción automática estadística

La traducción automática estadística (TAE) es un ejemplo de traducción automática basada en corpus. En este caso, el motor de traducción utiliza grandes volúmenes de corpus y textos paralelos, tanto bilingües como monolingües, además de un sistema estadístico que produce la traducción. Algunos ejemplos de este sistema son Moses (Koehn et al, 2012) o MTradumatica (Martín-Mor et al, 2017). El sistema Microsoft Translator (Microsoft, 2016) también utiliza un modelo de traducción automática estadística. En este caso, Microsoft ha estado recopilando textos durante más de diez años para alimentar el motor y producir así las traducciones. Aunque la API que ofrece en el momento en el que se realizó este estudio funcionaba con traducción automática estadística, Microsoft está trabajando ya con la traducción automática neuronal (Microsoft Research, 2018). Asimismo, Microsoft tiene a disposición del usuario un sitio web en el que se pueden efectuar pruebas sobre qué traducción, si la del motor de traducción automática estadística o la del neuronal considera el usuario que es mejor.

2.2.4 Traducción automática neuronal

La traducción automática neuronal se construye, en primer lugar, a partir de corpus lingüísticos, a los que, posteriormente, se les añade una red neuronal extremadamente compleja. Son la tecnología más vanguardista en traducción automática. Castilho et al (2017) lo explican así:

> Neural models involve building an end-to-end neural network that maps aligned bilingual texts which, given an input sentence X to be translated, is normally trained to maximise the probability of a target sequence Y without additional external linguistic information.

Sin embargo, a pesar de los grandes avances que ha proporcionado la traducción automática neuronal, es importante tener en cuenta que todavía presenta puntos flacos. En primer lugar, crear un motor de traducción automática neuronal aún supone un coste muy elevado. Asimismo, resulta muy lento entrenarlo, no es totalmente eficaz a la hora de traducir terminología y, según la complejidad de la palabra, en ocasiones deja parte del segmento vacío (Wu et al, 2016).

Uno de los motores de traducción automática neuronal más conocidos en la actualidad es el sistema Google Neural Machine Translation. En el artículo publicado en el año 2016, los autores procuran explicar las ventajas de usar este motor, en comparación con el coste que puede suponer emplear otro motor de traducción automática neuronal. Algunos de los beneficios son:

> To accelerate the final translation speed, we employ low-precision arithmetic during inference computations. To improve handling of rare words, we divide words into a limited set of common sub-word units ("wordpieces") for both input and output. This method provides a good balance between the flexibility of "character"-delimited models and the efficiency of "word"-delimited models, naturally handles translation of rare words, and ultimately improves the overall accuracy of the system. (Wu et al, 2016)

2.3 Evaluación de la calidad de la traducción automática

Si nos centramos en un primer momento en la calidad de la traducción (sin especificar que sea automática o no), en el número de la revista Tradumàtica en el que se trata la calidad en traducción (Koby et al, 2016), se establecen varios puntos de confluencia: el primero de ellos consiste en la definición que la industria proporciona de «cliente»; el segundo se centra en que cualquier producto se puede entender como una traducción, de la que se espera cierto nivel de precisión y fluidez; el tercero focaliza la atención en la posibilidad de una calidad *perfecta* en cuanto a fluidez y precisión, puesto que este es un tema complicado, dado que la codificación entre dos idiomas no siempre es mecánica. El cuarto punto habla de la responsabilidad de los proveedores para con el cliente y el usuario final, mientras que el quinto punto trata de la convicción, tanto de la industria como de los estudios de traducción, de que se debe conseguir un método para medir la calidad de la traducción de la forma más objetiva posible, fijándose en los problemas que se deben corregir.

Es de crucial importancia determinar la calidad de un motor de traducción automática no solo para los usuarios (sean compradores o no), sino también para los investigadores y los propios desarrolladores. Según Martín-Mor et al (2016, 63):

> Actualment, la recerca em 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.

Existen varios métodos automáticos para evaluar la calidad de la traducción automática, ya sean escalas de medida automáticas o mediante la evaluación con la intervención humana.

2.3.1 Escalas de medida automáticas

Si bien existen varias escalas de medida automáticas, a continuación, se describen algunas de las más usadas en la actualidad, que miden los aciertos de la traducción automática:

- BLEU (BiLingual Evaluation Understudy): las métricas de BLEU (Papineni et al, 2002) son uno de los métodos más usados en la actualidad. Trabaja a partir de n-gramas y parte tanto de la traducción automática como de una o más traducciones humanas de referencia. Se define de la siguiente forma: «The principle of this method is to calculate the degree of similarity between candidate (machine) translation and one or more reference translations based on the particular n-gram precision». (Chéragui, 2010). Otra de las definiciones la encontramos en Chunyu y Tak-ming (2015): «It is based on counting the number of n-grams, namely sequences of consecutive word(s) of varying length, co-occurring in an MT output and in one or more versions of corresponding reference, usually each in the form of a sentence».
- NIST: se diseñó en el año 2005 por el Nationale Institute of Standards para mejorar el método BLEU. Wolk y Koržinek (2017) lo definen así:

> The NIST metric was designed to improve BLEU by rewarding the translation of infrequently used words. This was intended to further prevent inflation of SMT evaluation scores by focusing on common words and high confidence translations. As a result, the NIST metric uses heavier

> weights for rarer words. The final NIST score is calculated using the arithmetic mean of the n-gram matches between SMT and reference translations. In addition, a smaller brevity penalty is used for smaller variations in phrase lengths.

- METEOR: al igual que NIST, surgió por primera vez en el año 2005; tiene en cuenta aspectos que son indirectos en BLEU: Chunyu y Tak-ming lo definen de la siguiente forma: «To maximize the possibility of matching, it uses three word-mapping criteria: (1) exact character sequences, (2) identical stem forms of word, and (3) synonyms» (2015, 228).

A diferencia de BLEU que, como se comentaba antes, se basa en número de n-gramas, METEOR se centra en el *recall*:

> Recall (the proportion of matched n-grams to total reference n-grams) is used directly in this metric. In addition, METEOR explicitly measures higher order n-grams, considers word-to-word matches, and applies arithmetic averaging for a final score. Best matches against multiple reference translations can also be used. (Wolk y Koržinek, 2017, 3).

Otros métodos de evaluar la traducción automática son los sistemas de Word Error Rate (WER) y Translation Error Rate (TER), que en este caso, a diferencia de las escalas anteriores, miden los errores de la traducción automática:

- Word Error Rate (WER): surgió en el año 2007 de la mano de Popovic y Ney y se basa en comparar la traducción automática a nivel de palabra con la traducción de referencia. El problema que se presenta con este método es que una única palabra puede tener más de una traducción válida (Martín-Mor et al, 2016, 64).
- Translation Error Rate (TER): propuesta por Snover en el año 2006, se define según la distancia de edición, es decir, el número mínimo de ediciones necesarias para modificar una traducción de referencia. «TER measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation» (Snover et al 2006).

Es necesario mencionar que un método que falta en esta lista es el Position-independent word Error Rate (PER) propuesto por Tillman en 1997. Este compara

palabras de traducción automática con las de la traducción de referencia sin tener en cuenta la posición en la frase. Esta es la mayor desventaja de este método, que ya está prácticamente obsoleto.

Al trabajar con el sistema de evaluación de traducción automática en bruto, se debe tener en cuenta su funcionamiento, así como otros factores fundamentales. Por ejemplo, para leer correctamente los resultados referentes a los aciertos de la traducción automática, el sistema debería proporcionar un valor entre 0 y 1 que se pudiera interpretar adecuadamente en términos de calidad. Así, si el valor está más cercano a 1 sería un sistema de mayor calidad, mientras que si está más cercano a 0 sería de menos (Martín-Mor et al, 2016, 74).

De la misma forma, es imprescindible tener presente que existen otros factores que impiden establecer una correlación totalmente fiable entre el valor proporcionado por la herramienta de evaluación y la calidad del propio sistema de traducción automática:

> Entre els factors més importants hi ha, d'una banda, el tipus de sistema de TA utilitzat. Atès que aquestes mètriques són de caràcter estadístic, les traduccions en brut obtingudes amb sistemes de TAE degudament entrenats acostumen a obtenir millor resultats que no pas les obtingudes amb sistemes de TABR. Tanmateix, alguns errors de traducció provinents de la TABR són més fàcilment editables que els provinents de la TAE, i, per tant, són de més qualitat com punt de partida per a la postedició. D'altra banda, un altre dels factor més rellevant que suposa un inconvenient per a l'estimació de la qualitat de la traducció en brut és que hi acostuma a haver una correlació entre el valor de mètriques com BLEU i la valoració humana de la qualitat d'una traducció en brut quan el sistema de TA no és gaire bo, però aquesta correlació deixa de ser tan constatable a mesura que el sistema de TA millora (Martín-Mor et al, 2014).

2.3.2 Intervención humana en la evaluación de la traducción automática

Como se mencionaba con anterioridad, existen distintas formas de evaluar la traducción automática que requiere la participación humana. Entre estas, se encuentran las siguientes:

- *Scoring-based human evaluation*: cuando se usa este método, se pide a los evaluadores que puntúen cada sistema por oración. La puntuación media es la puntuación final del sistema. Las métricas más usadas son precisión y fluidez. La enciclopedia Routledge de traducción define la precisión como «*how much meaning of the source sentence is conveyed in the target sentence*», mientras que la fluidez mide «what degree the target sentence is smooth idiomatically and grammatically». Estos baremos se verán más adelante con la herramienta de DQF de TAUS.
- *Ranking-based human evaluation*: los evaluadores deben establecer un ranking de los resultados de la misma frase a partir de uno o más sistemas. Una vez más, esta es una de las pruebas que se puede realizar a través de la plataforma de TAUS.
- *Post-edit-based human evaluation*: los evaluadores tienen que poseditar los resultados de traducción automática de cada segmento.
- *Human translation edit rate* (HTER): se calcula para cada sistema a través de la valoración de los resultados del sistema y sus correspondencias de posedición.

2.3.3 Normas y modelos para la evaluación de la calidad

A la hora de evaluar la calidad, es necesario mencionar que existen tanto normas como instituciones que crean sistemas para establecer criterios que seguir. A pesar de ello, resulta muy complicado establecer qué es la calidad (Fields et al, 2014 y Koby et al, 2014). Entre las normas más esenciales, podemos encontrar la de la European Union of Association of Translation Companies (EUACT), la EN-15038, del año 2006, aprobada por el Comité Europeo de Normalización (CEN). Por otra parte, en el año 2012 aparece la primera norma ISO (International Organization for Sandardization) centrada en los servicios de traducción (ISO/TS 116669:2012, Translation projects – General guidance), que sigue vigente en la actualidad. Ahora mismo se están elaborando normas específicas para la traducción automática y la posedición (Martín-Mor et al, 2016, 118).

A día de hoy, cuando se habla de calidad, es esencial mencionar el Multidimensional Quality Metrics (MQM) (Lommel et al, 2014). Surgió a raíz del proyecto QTLaunchPad, financiado por la Unión Europea, y se basa en el LISA QA Model de LISA (Localization Industry Standards Association) en 1995 (Martínez Mateo, 2014).

En la actualidad, MQM engloba las siguientes dimensiones, en las que se encuentran los diferentes tipos de errores. Las dimensiones de los tipos de errores son ocho que, a su vez, se subdividen en distintas categorías, conformando así una red amplia:

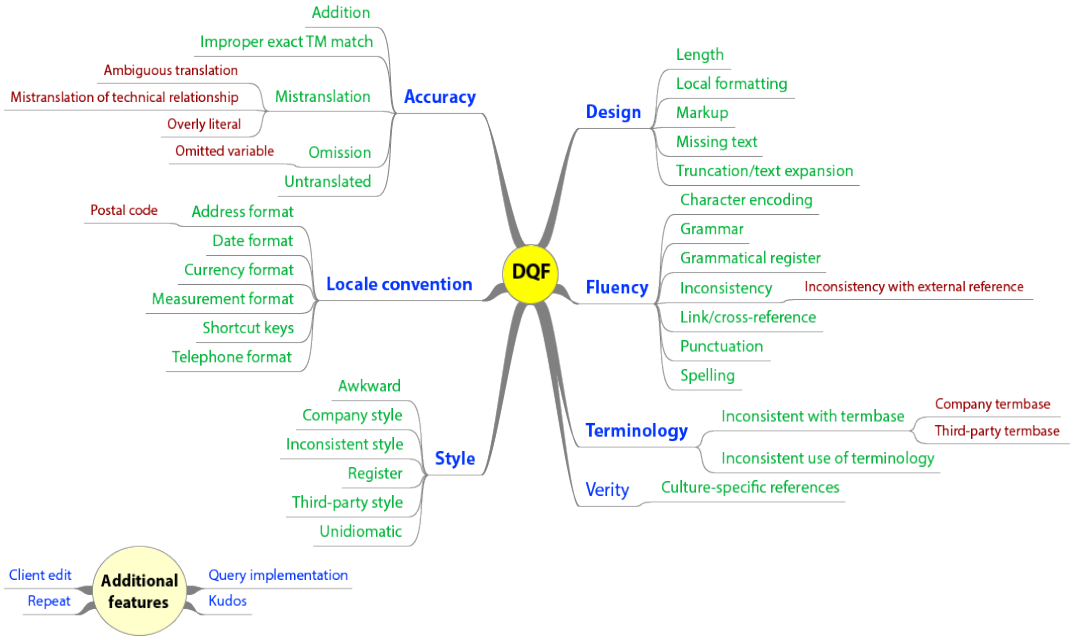


Figura 2: esquema de errores de DQF

En este trabajo se usarán dos de las clasificaciones del modelo de MQM para evaluar los resultados de las pruebas de calidad que conforman parte del marco metodológico de este proyecto, a saber, la fluidez y la precisión. En lo que se refiere a las pruebas, se realizaron mediante el Dynamic Quality Framework en el TAUS Dashboard. TAUS (Translation Automation User Society) es una asociación industrial que se creó en el año 2005 siguiendo la estela de otras asociaciones, como, por ejemplo, GALA. Fue fundada por grandes empresas de la industria de la localización. Con el

tiempo, la organización evolucionó hasta convertirse en una plataforma de colaboración centrada en la industria de la traducción global. Ha realizado grandes aportaciones a esta, entre las que destacan a día de hoy la Data Cloud y el Dynamic Quality Framework (DQF), así como el TAUS Quality Dashboard, en el que profundizaremos más adelante.

Si bien MQM y TAUS estaban concebidas en un principio como herramientas opuestas, con el tiempo se ha llegado a la conclusión de que se complementan mutuamente:

> As of late 2014, the two initiatives are in contact and working to harmonize their efforts. A close examination reveals them to be largely complementary. MQM provides a way to describe arbitrary metrics in a standardized fashion but does not provide guidance on the interpretation of the results. The DQF, by contrast, does not seek to describe all possible translation quality metrics but does provide guidance on interpreting quality evaluations for specific scenarios. Part of the ongoing work is to bring the two frameworks together such that any DQF metric can be described in MQM. This harmonization is planned as a key aspect in a forthcoming EU-funded project. (Lommel et al, 2014).

Los investigadores han usado las herramientas de TAUS, como el DQF, en los últimos tiempos para estimar la calidad de la traducción automática (O'Brien, 2012 y Görog, 2014). Desarrollado en 2011, el DQF es un intento de estandarizar la evaluación de la calidad, y se basa en lo siguiente:

> Quality in DQF is considered dynamic as translation quality requirements change depending on the content type, the purpose of the content and its audience. The Framework provides a commonly agreed approach to select the most appropriate translation quality evaluation model(s) and metrics depending on specific quality requirements. The underlying process, technology and resources affect the choice of the quality evaluation model. The Framework is underpinned by the recognition that quality is when the customer is satisfied.

2.4 Posedición

La ISO 18587:2017, que proporciona los requisitos para el proceso de la posedición de la traducción automática por humanos y las competencias de los traductores, define la posedición como «edit and correct machine translation output»; el *machine translation output*, por su parte, es el resultado de la traducción automática. Así, la traducción automática es la traducción automatizada de texto de un idioma natural a otro utilizando una aplicación de ordenador.

Asimismo, cabe destacar la información publicada en un informe de TAUS publicado en el año 2010, en la que la posedición sería «the process of improving a machine-generated translation with a minimum of manual labor». A la hora de poseditar, TAUS recomienda una serie de directrices que se deben tener en cuenta para obtener buenos resultados. Entre estas se encuentran, por ejemplo, asegurar la calidad del texto de partida o entrenar a los traductores con anterioridad.

2.4.1 Flujo de trabajo (TA + PE)

Es necesario tener en cuenta que el uso de la traducción automática implica un flujo de trabajo determinado. Por norma general, el flujo de trabajo de un proyecto de traducción consiste en tres grandes fases: obtención, gestión de proyecto y entrega. Dentro de la gestión de proyecto se pueden distinguir las etapas de análisis, preparación (estas dos son fases de preproducción), traducción inicial y posproducción.

Así, cuando se emplea la traducción automática, se parte de un texto que se conoce como traducción en bruto (*raw machine translation*, en inglés). Este sería un primer borrador del texto con el que se trabajará posteriormente.

Antes de preparar el texto para poseditarlo, se puede incluir una fase que consiste en la preedición del texto mediante lenguaje controlado. Cuando se emplea lenguaje controlado, al igual que en la preedición, se observan los mayores problemas del texto de origen para el motor de traducción automática (como pueden ser expresiones complejas o casos de ambigüedad) y se tratan previamente, antes de preparar el texto original para que sea una traducción automática en bruto.

Tras estas fases, se pasa a la fase de posedición. La enciclopedia Routledge de traducción define la posedición de la siguiente forma:

> Post-editing involves a human editor revising an MT output up to an acceptable level of quality. Quality of MT output is assumed to have an inverse correlation with the amount of effort needed for the revision. In this way MT is assessed as a means of raising translators' productivity in terms of the cost-effectiveness of post-editing its output as a usable initial draft of translation. In the worst case, this draft may take a translator even longer to post-edit than to translate its source text from scratch. (Chan, 2014).

Según la intención, el propósito de trabajo de la posedición y su visibilidad, la ISO/DIS 18587/2017 determina que se pueden distinguir dos tipos de posedición, a saber: *light post-editing* o posedición parcial y *full post-editing* o *high quality post-editing*, o posedición total[2]. Por una parte, la posedición parcial consiste en realizar solo el número de ediciones imprescindible para generar un texto que sea comprensible: «This involves taking the raw MT output and performing as few modifications as possible to the text in order to make the translation understandable, factually accurate, and grammatically correct» (Densmer, 2014). En este tipo de posedición, se busca que la traducción sea precisa, aunque pueda no ser tan fluida como una traducción humana (Martín-Mor et al, 2016, 70).

Por otra parte, la posedición de alta calidad consiste en producir un texto final con el que se busca el mismo estándar de calidad que una traducción humana (también conocido como *human quality*): «After this edit, the translation should read as if written in the target language» (Densmer, 2014).

Es necesario mencionar que existe un tipo de traducción automática distinto, conocida como traducción *desatendida* (Sánchez-Gijón, 2016). En estos casos, se publica la traducción en bruto antes de poseditarla, ya sea porque se quiere tener la versión disponible de forma inmediata o por ahorrar en gastos de posedición. En

[2] La primera vez que se hizo esta distinción fue en los años 80, de la mano de Loffler y Laurian (1986) en las que se distinguía entre posedición rápida y convencional.

muchas ocasiones, cuando esto sucede, se le suele presentar al usuario final un cuestionario para que valore la calidad de la traducción. De esta forma, se puede determinar si es necesario realizar o no una posedición posterior.

2.4.2 Tipos de posedición

Según los recursos técnicos, se pueden distinguir dos tipos de posedición:

- Posedición clásica: debido a sus características, es la más sencilla; al poseditor se le presenta el texto original, en el idioma de origen, y la traducción en bruto, generalmente ya de forma segmentada. Se puede realizar directamente en un entorno de edición o con cualquier procesador de texto u hojas de cálculo (Martín-Mor et al, 2016, 70).
- Posedición integrada: para llevarla a cabo, es imprescindible contar con una herramienta TAO. En este caso, el poseditor recibe tanto el texto original como la traducción en bruto y la memoria de traducción (*translation memory*, TM, por sus siglas en inglés). Así, el poseditor puede partir directamente de la traducción en bruto o de una traducción que figure en la TM y que hubiera sido validada con anterioridad. La posedición integrada puede ser:
    - Posedición en vivo: en este caso, se conecta la herramienta TAO con un sistema de TA. De esta forma, el poseditor recibe una propuesta de TA en los casos en los que no hay ninguna traducción disponible en la TM.
    - Posedición *in vitro*: ya sea por motivos de conectividad o de confidencialidad, en este caso no se conecta directamente la herramienta TAO con el sistema de TA, sino que se vuelca primero en una memoria de traducción. Sigue siendo imprescindible trabajar con una herramienta TAO, dado que es la conexión con esta lo que permite que, al detectar el segmento vacío, se baje automáticamente la traducción en bruto.

Martín-Mor et al engloban todas estas características en cuatro grandes dimensiones al trabajar en la fase de la posedición: cómo, cuándo, quién y con qué propósito (2016, 66).

2.4.3 Modos de integración de la TA en el flujo de traducción

A la hora de hablar de traducción automática, es indispensable no perder de vista los conceptos de *Fully Automatic High Quality Machine Translation* y *Human-Aided Machine Translation* (Hutchins, 1986).

- *Fully automatic high quality machine translation* (FAHQT): la primera vez que apareció este concepto fue en los años cincuenta, cuando Yehoshua Bar-Hillel lo usó por primera vez. Bar-Hillel defendía que tratar de alcanzar la FAHQT no solo era poco realista, sino imposible. Según él, no habría ningún ordenador capaz de replicar las habilidades y el conocimiento humano, lo que hoy se conoce como *real world knowledge*. Sus creencias se incluyeron en el informe ALPAC. Aunque hoy en día se ha mejorado mucho en este campo, en especial gracias a la ayuda de la Inteligencia Artificial, todavía falta tiempo para alcanzar un sistema de traducción automática que proporcione una calidad total sin que sea necesaria la intervención humana en algún punto.
- *Human-aided machine translation* (HAMT): se trata de que el sistema toma todo el peso de la situación, mientras que la intervención humana se ve reducida a momentos puntuales, ya sea durante el proceso interno o externo. Durante el proceso interno se hace referencia únicamente a la forma interactiva, en el proceso de análisis, transferencia y generación, cuando el sistema necesita ayuda para tomar decisiones léxicas y resolver las ambigüedades. Con proceso externo nos referimos al proceso de predición[3] y posedición.
- *Machine-aided human translation* (MAHT): en este caso se hace referencia a cualquier herramienta computacional (ya sea un sistema o un programa) de ayuda lingüística. Así, en esta categoría encontraríamos programas que comprueban la ortografía, la gramática o el estilo de la traducción. Por otra

[3] Hutchins & Sommers (1986) definen la preedición como el proceso que consiste en comprobar si el texto de origen presenta algún problema para el sistema de TA y solucionarlos antes de pasar el texto por el motor.

parte, también figurarían aquí las referencias en línea como diccionarios (monolingües o bilingües), tesauros, enciclopedias, glosarios, etc. En la actualidad, muchos de estos procesos se engloban en los sistemas de gestión y edición de traducciones (SGET) (Martín-Mor et al, 2015, 96).

2.4.4 La traducción automática en el flujo profesional

Si bien es cierto que el informe ALPAC marcó un antes y un después en la historia de la traducción automática, a día de hoy, la percepción en el mercado es que todavía queda mucho por hacer para llegar a alcanzar el ideal de FAHQT. No obstante, los avances en tecnología y el desarrollo de la inteligencia artificial están haciendo que los sistemas de traducción automática estén cambiando y se desarrollen rápidamente. Asimismo, las necesidades del sector evolucionan constantemente y se está volviendo imperioso traducir un número mayor de palabras en el menor tiempo posible, así como abaratar los costes. Por tanto, el uso de la traducción automática se está volviendo indispensable, siempre y cuando se pueda establecer un estándar de calidad y que el coste (ya sea económico o de esfuerzo) se vea cubierto por las ventajas y los beneficios.

Con el paso del tiempo, la traducción automática se ha incluido poco a poco en el flujo de trabajo de la traducción. Así, en el artículo de Presas, Cid-Leal y Torres-Hostench (2016) se puede observar que las LSP de una región concreta ya han incorporado la traducción automática y la posedición a sus rutinas. En lo que se refiere a posedición, SánchezGijón (2016) busca definir las competencias del poseditor, así como ampliar la concepción de la posedición, desde una más reducida (el traductor/poseditor se limita a validar los segmentos) a una más amplia (el traductor/poseditor domina todo el flujo de trabajo relacionado con la traducción automática), fundamentando esta distinción en la competencia tecnológica. En esta línea, se busca otorgar una dimensión tecnológica al perfil del traductor, complementando la proporcionada por Rico y Torrejón (2016).

En cuanto a la evolución de las herramientas de traducción, esta queda patente en los cambios que han experimentado los sistemas de traducción automática. Si bien en

un principio se comenzó trabajando con sistemas basados en corpus paralelos (sistemas de traducción estadística) y aquellos basados en reglas, a día de hoy, con el uso de la inteligencia artificial, se están desarrollando sistemas de traducción automática neuronal cada vez mejores y más avanzados.

A pesar de estos avances, es fundamental que se siga perfeccionando la búsqueda de mejorar la calidad de la traducción bruta obtenida, así como aumentar la productividad y reducir el esfuerzo de posedición. Esto lo trata Guerberof (2009) en su estudio. En él, Guerberof realiza una prueba con nueve traductores profesionales a través de una prueba de posedición en línea. Debían traducir segmentos nuevos, de traducción automática estadística y procedentes de una memoria de traducción (coincidencias parciales entre el 80 % y el 90 %) sin saber el origen de cada segmento. Los resultados referentes a la productividad indicaron que los traductores eran más rápidos poseditando que traduciendo los segmentos de cero e incluso editando las coincidencias parciales de la memoria de traducción. En cuanto a los errores encontrados evaluando la calidad, más de la mitad procedían de los segmentos de la memoria de traducción, mientras que el 27 % era de la traducción automática y el 21 % en los segmentos nuevos. En cuanto a los errores, el 44 % eran de precisión. No se puede establecer cuántos fueron los errores de fluidez, ya que para este estudio se utilizó la escala de LISA, en la que no se incluye la fluidez como una categoría de errores.

En otro orden de cosas, aunque los sistemas de traducción automática neuronal sean una tecnología de vanguardia, todavía surgen problemas a la hora de trabajar con ellos. Uno de los más importantes es, por ejemplo, el tratamiento de la terminología (Luong et al, 2016). En este artículo, se presenta la problemática de traducir terminología con sistemas de traducción automática neuronal.

Asimismo, cabe destacar el estudio llevado a cabo por Bentivogli et al (2016), en el que se establece una comparativa entre los resultados obtenidos con un motor de traducción automática neuronal y un motor de traducción automática estadística basado en frases. El objetivo de este estudio era conocer las fortalezas de los sistemas de

traducción automática neuronal y las debilidades del de estadística basado en frases. Para ello, se analizaron errores morfológicos, léxicos y de ordenación de palabras.

En cuanto a la percepción de los traductores, en este caso sobre la posedición, Guerberof realizó un estudio en el año 2013 en el que consultaba a un grupo de traductores mediante una encuesta qué opinaban de la posedición. A través de una serie de preguntas, se concluyó que la productividad de los traductores se mantenía igual a lo largo del tiempo (esto lo opinaba el 45 % de los traductores), mientras que el 40 % creía que la productividad había aumentado con el paso del tiempo. Ningún traductor respondió que su productividad había decrecido. Además, el 55 % de los traductores consideraba que la experiencia que se adquiría poseditando aumentaba las capacidades de detectar los errores, mientras que el 30 % creía que tenía la misma capacidad a la hora de detectarlos que al principio. En estos dos casos, un 15 % consideraba que no sabía cómo responder.

En el artículo de Läubli et al (2013), se explica la importancia de la posedición en un entorno conocido para los traductores. En el estudio, se exponía la necesidad de realizar las distintas pruebas para recabar los datos en un entorno controlado, mientras que en el caso que se presentaba se explicaba la relevancia de utilizar una herramienta TAO que los traductores conocieran. Así, para llevar a cabo las pruebas de este estudio se empleó Across, una herramienta TAO que los traductores ya conocían. En este artículo, se realizó un experimento en el que se usaba tanto traducción automática como concordancias provenientes de una memoria de traducción. Finalmente, los resultados indicaron que la posedición reducía el tiempo invertido en un 15-20 %, un tiempo inferior al conseguido en otros estudios, en los que el ahorro se situaba estaba en un 40 % (Plitt y Masselot, 2010, y Sousa et al, 2011).

Siguiendo con esta línea, en el artículo de Moorkens y O'Brien (2017), se busca entender las necesidades que tienen los poseditores al llevar a cabo esta tarea. Los autores encuestaron a 231 participantes y entrevistaron a diez de ellos, centrándose en la posedición de traducción automática. No obstante, los resultados destacaron la insatisfacción continuada de los traductores con las herramientas TAO. A pesar de los

constantes avances en las herramientas TAO en los últimos años, los traductores tienen la percepción de que falta una evolución con las interfaces ya existentes.

En otro orden de cosas, en un apartado del marco metodológico, concretamente del diseño metodológico, se mencionará la importancia de la longitud de los segmentos a la hora de poseditar. Tal y como han demostrado estudios recientes, la calidad de la traducción neuronal baja considerablemente a medida que aumenta la longitud de los segmentos. Xuang y Xiong (2016) establecen en su estudio un método mediante el cual se acortan los segmentos más largos de forma automática en distintas cláusulas. Estas cláusulas se traducen con el motor de forma automática y se vuelven a ordenar sin modificar su orden. Tras realizar diversas pruebas, los resultados demuestran que, mediante este sistema, los puntos BLEU aumentan 2,94 con respecto a la línea de media al utilizar esta segmentación en segmentos de más de 30 palabras, y 5,43 en los segmentos de más de 40 palabras.

3 Marco metodológico

Tal y como se ha visto en el apartado de Objetivos presentado en la introducción de este trabajo, con este estudio se pretende determinar si existen diferencias entre la traducción automática neuronal y la traducción automática estadística. En concreto, se buscan aquellas relacionadas con la percepción de los traductores de los resultados de estos dos tipos de motores y la productividad, entendida como tiempo de posedición y distancia de edición, al poseditar los textos. Así, se parte de la hipótesis de que los resultados producidos por un sistema de traducción automática neuronal se perciben como de mejor calidad que aquellos resultantes de un sistema de traducción automática estadística, y que con la primera se obtienen mejores resultados en términos de productividad (se posedita más rápido con la traducción automática neuronal que con la estadística) y de distancia de ediciones (se edita menos al usar traducción automática neuronal que estadística).

Precisamente debido a que lo que se quería obtener era la percepción real de los traductores, este estudio se ha realizado con un diseño empírico-experimental. Para ello se han empleado dos tipos de textos distintos, así como varios participantes, todos ellos traductores (nueve en la primera fase del proyecto y seis en la segunda). El motivo de emplear dos textos distintos radica en querer cruzar los resultados para obtener la mayor cantidad de información posible.

Tras esta breve introducción, este apartado se estructurará, en un primer momento, en la presentación de las pruebas realizadas y la preparación efectuada para llevarlas a cabo. Posteriormente, se presentan los dos diseños metodológicos de este proyecto. Dado que en este trabajo se pretenden observar dos objetos de estudio distintos (por un lado, la percepción, y, por otro, la productividad), se han usado dos diseños metodológicos distintos. Si bien se ha trabajado con la calidad de los segmentos de traducción automática en bruto, las pruebas relacionadas con este objeto sirven como instrumento para calcular los resultados a la hora de hallar los resultados relacionados con la productividad y la distancia de edición.

Así, el diseño empírico-experimental de este proyecto se divide en dos partes claramente diferenciadas. Dado que son dos los objetos de estudio, a saber, la percepción y la productividad de los traductores al usar estos dos sistemas de traducción automática, este proyecto presenta dos diseños metodológicos diferenciados. El tipo de investigación realizada en todo momento es cuantitativa, dado que únicamente se recopilaron los datos a partir de la herramienta empleada para las pruebas (esto es, los proyectos del Dynamic Quality Framework de TAUS). Las variables controladas de este proyecto son los sujetos, los textos y las instrucciones breves sobre cómo se debe poseditar, tal y como se verá a continuación.

En primer lugar, los traductores llevaron a cabo una prueba de Quick Evaluation en la herramienta de DQF de TAUS. Posteriormente, un grupo más reducido de los mismos traductores realizó una prueba de productividad en la misma herramienta de DQF. Al mismo tiempo que esto ocurría, se efectuaron varias pruebas de calidad, en las que se evaluó tanto la fluidez como la precisión de los dos tipos de motores. El motivo por el cual se escogió este orden es que permite recabar en primer lugar toda la información necesaria para determinar la percepción de los traductores. Para finalizar, al realizar las pruebas de productividad y obtener los resultados correspondientes, se puede cotejar esta información con la recopilada en la prueba inicial de evaluación y, así, contrastarla.

3.1 Descripción de las pruebas

Todas las pruebas realizadas se llevaron a cabo mediante el DQF de TAUS. Para ello, fue necesario crear los tres tipos de prueba que permitían llegar a las conclusiones que se presentaban al inicio de la investigación como objetivos, a saber:

- *MT Ranking*: el objetivo que se pretende cumplir mediante esta prueba es determinar la percepción de los traductores sobre la calidad de los dos motores. Desde TAUS (Görög, 2017), la prueba se define de la siguiente forma:

  > The Comparison Task helps users select MT engines or human translators based on the quality of the output. DQF limits the number of sources you can compare to three. Shared experience at TAUS member companies has

> shown that an evaluator's ability to make robust judgments is impaired if he or she has to score more than 3 options segment-by-segment. After the translation files are uploaded, evaluators are invited to compare the translated segments and to give a ranking.

En este caso, la prueba la llevaron a cabo nueve traductores con los dos motores mencionados para los textos de marketing y de documentación. Para ello se utilizó una única memoria de traducción en formato TMX que contenía todos los segmentos de todas las variantes (tanto de marketing como de documentación, así como de traducción automática neuronal y estadística).

- *Quality Evaluation* (prueba de calidad): con esta prueba se pretende valorar la calidad de los dos motores de traducción. Si bien la prueba de DQF permite valorar todos los errores del MQM, para este estudio nos centramos únicamente en la fluidez (*fluency*) y la precisión (*accuracy*), que se categorizan de forma numérica del 1 al 4 según sean menos o más fluidos o precisos.

Para esta prueba, es esencial definir primero los valores que se van a tener en cuenta a la hora de realizarla. DQF utiliza las definiciones del Linguistic Data Consortium, tal y como aparece en el artículo publicado por A. Görög en el año 2014, en los casos de fluidez y precisión:

- Adequacy: "How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation."
- Fluency: "To what extent the translation is "one that 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."

De esta forma, el evaluador comprende cuáles son las características y requisitos que debe tener en cuenta.

- *Productivity test* (prueba de productividad): con esta prueba se pretende determinar el tiempo que emplea cada traductor a la hora de poseditar los distintos textos y motores. La realizan los traductores directamente en la

plataforma de DQF, tras recibir un enlace por correo electrónico. En la plataforma de DQF, el traductor puede ver el segmento original, el segmento de destino con la traducción automática en bruto. Görög (2014, 452) nos da la siguiente información:

> Post-editing productivity testing is becoming one of the most practical ways of generating evaluation scores. In this evaluation, users can choose to either post-edit the entire MT-output or translate half of the segments from scratch and post-edit the other half. In the latter case, the DQF tool removes half the target side (MT output) segments from the uploaded file(s). In both cases, the system measures the edit distance and the time taken to complete the tasks. When assigning the task to users, you need to specify which of two types of post-editing is required (i.e. light or full).

3.2 Preparación de las pruebas

Para preparar las pruebas, fueron necesarios varios instrumentos. Algunos de ellos eran comunes a todas las pruebas, mientras que otros eran de uso exclusivo para la prueba en cuestión. En este apartado también se verán los sujetos que participaron en cada prueba.

3.3 Instrumentos empleados

Para llevar a cabo este proyecto se emplearon varios instrumentos, dependiendo de la fase y de los resultados que se querían obtener. Es necesario precisar que hay instrumentos que son comunes a todas las pruebas, mientras que otros son específicos para cada diseño.

3.3.1 Instrumentos en común

En primer lugar, se escogieron los textos para llevar a cabo las distintas pruebas. Con el objetivo de obtener unos resultados válidos en la actualidad y dado el contexto del mercado, se escogió, por un lado, un texto de un manual de usuario de un *smartwatch*, a través de un archivo PDF disponible para su descarga en internet. Por otro, se buscó un texto de marketing, con una temática similar a la del manual de usuario (en este caso, un texto sobre las características de un teléfono móvil). El motivo por el cual se han escogido dos tipos de textos distintos es para abarcar un mayor rango

de resultados. De esta forma, si fuera necesario, se podría observar si existe alguna diferencia según la tipología textual. En ningún momento se trabajó con unos textos diferentes de los aquí presentados, ni se intervino de forma alguna, salvo en la adaptación de los textos para las pruebas, como se verá a continuación. En cuanto a la razón de escoger un texto de marketing y otro de un manual de usuario, se debe a que son dos de los tipos de texto más comunes que recibe una agencia de traducción. Aunque la longitud de los textos era considerable, se acortaron a aproximadamente unas 2000 palabras, para que fuera más fácil gestionarlos y que, al mismo tiempo, la herramienta de DQF los aceptase, ya que esta tiene un mínimo válido de segmentos (250 segmentos).

Para trabajar con los textos, se crearon varios proyectos en la herramienta TAO SDL Trados Studio 2017. De esta forma se conseguía utilizar la herramienta TAO para trabajar de una forma sencilla con los documentos, así como para prepararlos para su uso en la posedición. Así, a partir de estos proyectos se crearon cuatro memorias de traducción, en un principio vacías, con el formato SDLTM. En estas memorias se volcaron posteriormente los resultados de la traducción automática, que provenía de dos API, generadas una en Microsoft Azure y otra en Google Cloud. Tras esto, se exportaron al formato de intercambio TMX.

3.3.2 Instrumentos para cada prueba

En este apartado se describen los instrumentos empleados según las pruebas realizadas por los participantes:

3.3.2.1 Memorias de traducción

Tal y como se ha visto con anterioridad, una de las primeras herramientas empleadas fue la herramienta TAO SDL Trados Studio 2017. Si bien esta herramienta está enfocada a la traducción mediante la creación de proyectos, en este caso se empleó únicamente para la primera fase del proyecto, esto es, para la preparación de los textos y la gestión de las memorias de traducción. De esta forma, se crearon las memorias de traducción correspondientes a cada texto y a cada tipo de motor (Marketing – TAE, Marketing – TAN, Documentación – TAE y Documentación – TAN) en el formato

propio de SDL (SDLTM) y se exportaron al formato TMX. Esto es debido a que los distintos proyectos creados en DQF para realizar las pruebas solo aceptaban archivos con formato CSV y TMX. Para facilitar la gestión de las memorias, se decidió trabajar únicamente con estas en este software, en lugar de hacerlo, por ejemplo, en una hoja de datos. En cuanto a las memorias de traducción se refiere, es necesario añadir que se creó una única memoria de traducción que albergaba todos los segmentos de las cuatro variantes posibles, dado que esta sería necesaria en una de las pruebas que realizarían los traductores.

Se debe precisar que la herramienta TAO SDL Trados Studio, en su versión de 2017, permite conectar directamente la API de Google, de la que hablaremos en el siguiente apartado, en la configuración de proyectos. No obstante, esto no sucede con la API de Microsoft Azure. Para solventar este problema, se instaló el complemento MT Enhanced, disponible de forma gratuita en línea desde la App Store de SDL, que permite usar servicios de traducción automática que no están directamente integrados en el software por defecto de SDL Trados Studio. Para instalarlo solamente fue necesario descargarlo desde la App Store. Esta instalación habilita una nueva opción en la configuración de proyectos de la herramienta TAO que permite añadir la API generada de Microsoft Azure.

Al margen del programa SDL Trados Studio, en su versión de 2017, se emplearon otros para trabajar correctamente con las memorias de traducción. Cabe mencionar que se usó el software libre Notepad++, un editor de texto, para unificar las memorias de traducción en formato TMX en una única memoria de traducción. Este paso será necesario para llevar a cabo la prueba de MT Ranking con el conjunto de nueve traductores. Para descargar y editar los PDF de los textos de trabajo, se usó la herramienta de trabajo Adobe Acrobat. Este programa permitió un primer análisis general de los textos, así como su edición posterior mediante la eliminación de las páginas determinadas.

3.3.2.2 API de traducción

Para crear las memorias de traducción en la herramienta TAO a la hora de preparar los distintos proyectos, previamente fue necesario crear las dos API que permitían acceder directamente desde la herramienta. Para ello se emplearon las plataformas de Google Cloud y de Microsoft Azure, en las que fue necesario crear una cuenta para cada plataforma y darse de alta como usuario mediante cuentas de correo electrónico (a saber, Gmail y Outlook). Las dos cuentas contaban con un periodo de prueba gratuito que permitía interactuar con muchas de las características de las dos plataformas, entre las que se encontraban las API de traducción. Tras crear los usuarios a través de las dos cuentas, se dieron de alta los servicios de traducción en las dos plataformas y se generaron las dos API correspondientes.

3.3.2.3 Pruebas del Dynamic Quality Dashboard de TAUS

Por último, la herramienta empleada para realizar las distintas pruebas con los traductores fueron distintos tipos de proyectos del Dynamic Quality Framework de TAUS. Como se ha visto con anterioridad en el apartado del marco teórico, el Dynamic Quality Dashboard es una plataforma creada por TAUS para evaluar la calidad y la productividad de una forma dinámica, así como para convertir los datos de los que ya se disponen en útiles para la industria. La justificación de utilizar el DQF de TAUS como parte de este diseño metodológico radica en que una única herramienta permitía valorar los distintos aspectos que se buscan obtener con esta investigación.

3.3.3 Sujetos

En total, nueve traductores participaron en la primera prueba realizada, la de MT Ranking, todos ellos con edades comprendidas entre los 25 y los 32 años, con estudios ya finalizados del Grado de Traducción e Interpretación o de Licenciatura en Traducción e Interpretación. Su experiencia profesional comprendía entre uno y tres años en el campo de la traducción, ya sea como traductores autónomos o trabajando en LSP o MLV como traductores o gestores de proyectos.

Es necesario destacar que ocho de los nueve traductores habían cursado algún máster relacionado con Traducción. Una vez más, todos los participantes recibieron un

correo electrónico con unas directrices breves acerca de las pruebas, así como un consentimiento informado en el que se les informaba de en qué consistía el proyecto y para qué se usarían los resultados obtenidos. Asimismo, se les avisaba de que la información recopilada no se usaría para ningún fin que no fuera académica. Cabe destacar que todos ellos conocían qué era la posedición y habían trabajado profesionalmente con ella, ya fuera poseditando algún proyecto en alguna de las combinaciones de idiomas con las que trabajan o como gestor de un proyecto de posedición.

Para estas pruebas de productividad, solamente se emplearon seis de los nueve traductores que participaron en la prueba de MT Ranking. Esto es debido a que, si hubiesen participado los nueve, se hubiera obtenido una cantidad de resultados demasiado alta y difícil de procesar.

3.3.3.1 Preparación de las pruebas

La preparación de las pruebas consistió en la selección y preparación de los textos y la preparación de las memorias. El resultado son dos textos y cuatro memorias, además de una memoria de traducción que contenía todos los segmentos de las cuatro memorias anteriores.

3.3.3.2 Preparación de los textos

En primer lugar, se escogieron los dos textos correspondientes. Tras su descarga, se analizaron mediante Adobe Acrobat y acortaron para adaptarlos a las pruebas con las que trabajarían más adelante los traductores. Para la fase de edición, se usó Microsoft Word, de forma que se pudiera comprobar el volumen de palabras de cada texto antes de comenzar a preparar las memorias.

3.3.3.3 Preparación de las memorias

Tras la preparación de los textos, se crearon varios proyectos de traducción en la herramienta SDL Trados Studio 2017 y se creó una memoria de traducción en formato TMX con los segmentos de *raw machine translation*. Para trabajar con comodidad, se crearon sendos proyectos de trabajo con el programa SDL Trados Studio 2017. Al

mismo tiempo, se abrió una cuenta con prueba gratuita en Google Cloud y una en Microsoft Azure para crear las dos API que permiten conectarlas a la herramienta TAO elegida.

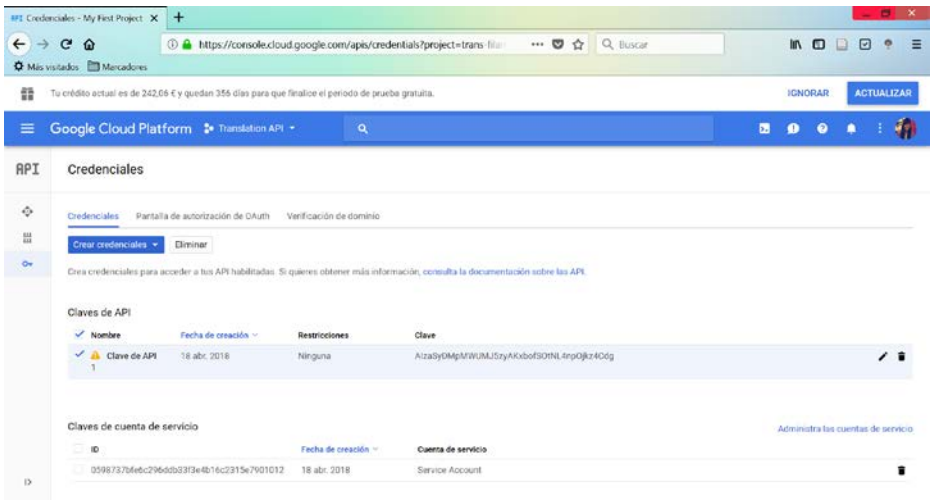


Figura 3: plataforma en línea de Google Cloud

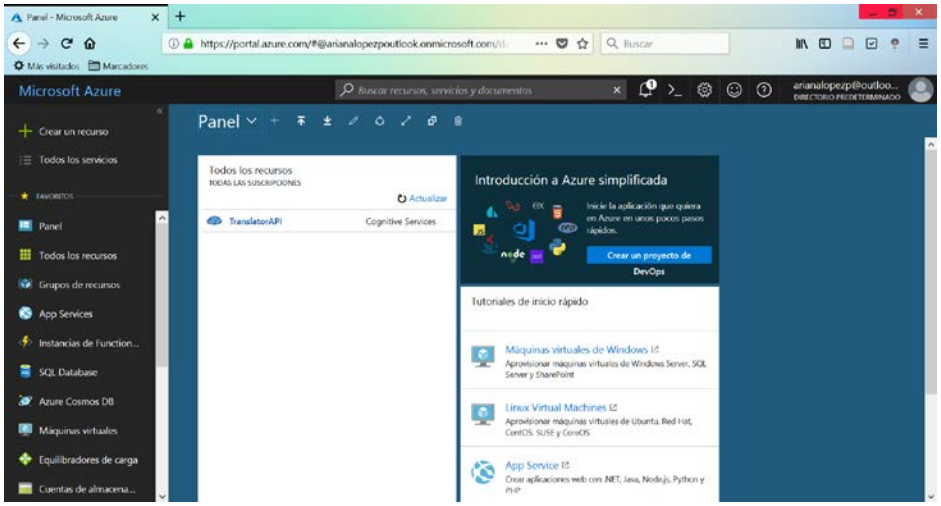


Figura 4: plataforma en línea de Microsoft Azure

En el caso de Microsoft Azure, para conectar la API a SDL Trados, fue necesario habilitar el complemento de SDL Trados MT Enhanced. Sin realizar este paso de forma previa, no es posible conectar la API de Microsoft Translator a la herramienta, tal y como se presenta en la siguiente imagen:

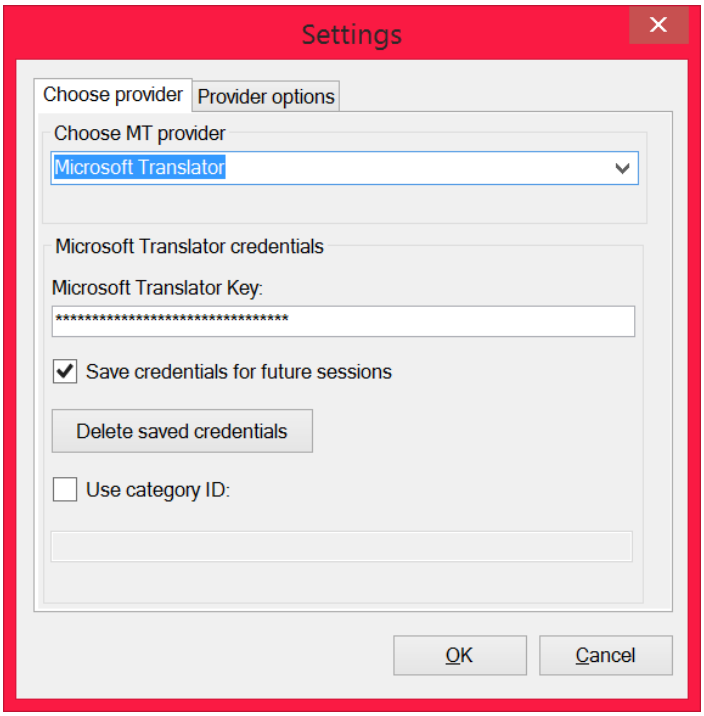


Figura 5: ventana de confguración del complemento en Sdl Trados Studio

Una vez permitido el acceso de las dos API en la herramienta y con los dos proyectos creados, se procede a la preparación de los archivos. Dado que será necesario disponer de cuatro memorias de traducción diferentes (Marketing – TAE, Marketing – TAN, Documentación – TAE y Documentación – TAN) para realizar todas las pruebas necesarias, se crearon dos proyectos, uno para el texto de marketing y otro para el de documentación, y se utilizaron los archivos SDLXLIFF de la combinación de origen (en-UK) para trabajar con ellos desde cero (con los segmentos vacíos, sin que contuviesen ninguna traducción automática) las veces que fueran necesarias. Posteriormente, mediante la tarea de «Procesamiento por lotes > Pretraducir», se pretradujeron los archivos tanto con la API de Google como con la de Microsoft. Tras

esto, y mediante la tarea por lotes «Llenar memorias de traducción del proyecto» se volcaron los segmentos en las memorias de traducción. Para finalizar, se exportaron las memorias de traducción propias de SDL Trados Studio (SDLTM) a un formato intercambiable (TMX), para trabajar con ellas correctamente en DQF. Así, las memorias tenían aproximadamente 2500 palabras cada una y cerca de 250 segmentos, el número mínimo requerido por DQF para realizar la prueba.

3.4 Prueba 1: MT Ranking

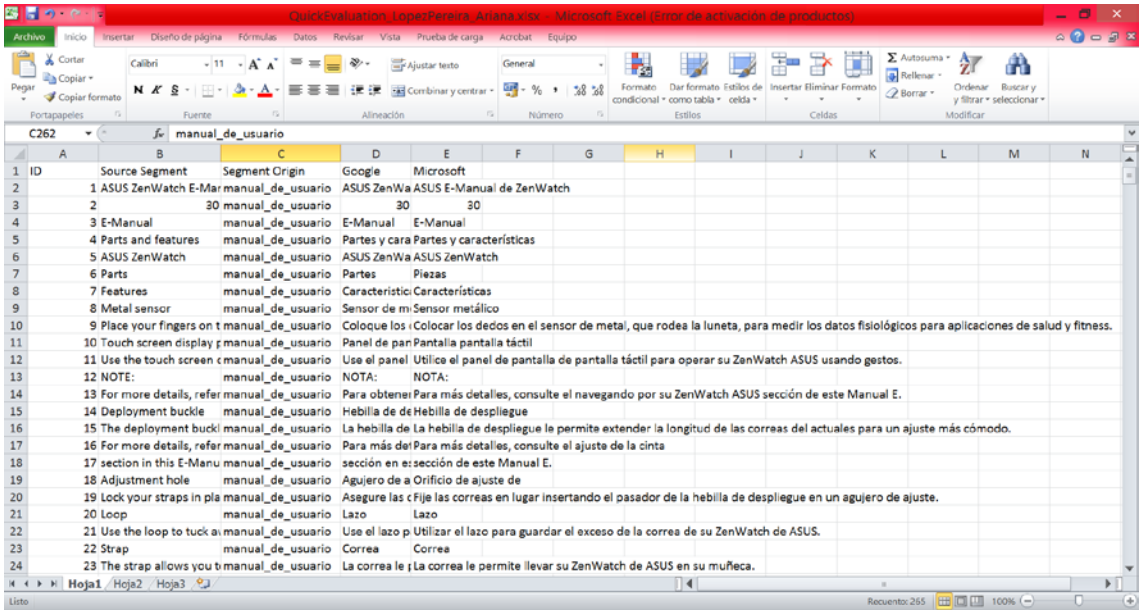
La primera fase de este proyecto consistió en realizar una prueba de MT Ranking, denominada así, en DQF. El objeto de estudio en este caso es la percepción de los traductores sobre estos dos sistemas de traducción. El método empleado es empírico-experimental.

Tal y como se ha visto con anterioridad, esta prueba consiste en construir una única memoria de traducción con los segmentos originales y los equivalentes correspondientes de los sistemas de traducción automática. Si bien es posible cotejar tres motores de traducción automática simultáneamente, para este proyecto se decidió utilizar solamente dos, dado que lo que se pretende cotejar es la percepción de los traductores en el uso de los motores de TAE y TAN. Al mismo tiempo, esta tarea permite llevar a cabo la evaluación de varios motores de traducción automática sin que los traductores (o evaluadores) sepan en ningún momento qué motor es el que están evaluando. De esta forma, los traductores no tienen prejuicios a la hora de decidir entre los segmentos que se les presentan. Es necesario mencionar que se valoró la posibilidad de emplear un cuestionario para determinar previamente la percepción de los traductores sobre los motores de traducción automática. No obstante, finalmente esta posibilidad se descartó, dado que la recopilación de esta información probablemente proporcionaría resultados sesgados por las propias opiniones y prejuicios de los traductores sobre los sistemas debido a su experiencia previa en la profesión. A través de una prueba a ciegas, se puede obtener la misma información sin que las opiniones previas de los traductores afecten al estudio.

El funcionamiento de la prueba consiste en que a los evaluadores se les presenta una única prueba de evaluación a partir de las memorias de traducción creadas en un primer lugar. Para llevarla a cabo, es necesario proporcionarle un archivo CSV a la herramienta de DQF con la siguiente estructura en la primera fila:

- ID
- Source segment
- Segment origin
- Motor de TA 1
- Motor de TA 2

Posteriormente, hay que volcar las memorias de traducción que se habían preparado al comienzo del proyecto a través de la herramienta TAO, en este caso, SDL Trados Studio. El resultado que se obtiene es el siguiente archivo:



ID	Source Segment	Segment Origin	Google	Microsoft
1	ASUS ZenWatch E-Mar	manual_de_usuario	ASUS ZenWa	ASUS E-Manual de ZenWatch
2	30	manual_de_usuario	30	30
3	E-Manual	manual_de_usuario	E-Manual	E-Manual
4	Parts and features	manual_de_usuario	Partes y cara	Partes y características
5	ASUS ZenWatch	manual_de_usuario	ASUS ZenWa	ASUS ZenWatch
6	Parts	manual_de_usuario	Partes	Piezas
7	Features	manual_de_usuario	Caracteristic	Características
8	Metal sensor	manual_de_usuario	Sensor de m	Sensor metálico
9	Place your fingers on t	manual_de_usuario	Coloque los	Colocar los dedos en el sensor de metal, que rodea la luneta, para medir los datos fisiológicos para aplicaciones de salud y fitness.
10	Touch screen display p	manual_de_usuario	Panel de pan	Pantalla pantalla táctil
11	Use the touch screen c	manual_de_usuario	Use el panel	Utilice el panel de pantalla de pantalla táctil para operar su ZenWatch ASUS usando gestos.
12	NOTE:	manual_de_usuario	NOTA:	NOTA:
13	For more details, refer	manual_de_usuario	Para obtener	Para más detalles, consulte el navegando por su ZenWatch ASUS sección de este Manual E.
14	Deployment buckle	manual_de_usuario	Hebilla de de	Hebilla de despliegue
15	The deployment buckl	manual_de_usuario	La hebilla de	La hebilla de despliegue le permite extender la longitud de las correas del actuales para un ajuste más cómodo.
16	For more details, refer	manual_de_usuario	Para más det	Para más detalles, consulte el ajuste de la cinta
17	section in this E-Manu	manual_de_usuario	sección en es	sección de este Manual E.
18	Adjustment hole	manual_de_usuario	Agujero de a	Orificio de ajuste de
19	Lock your straps in pla	manual_de_usuario	Asegure las c	Fije las correas en lugar insertando el pasador de la hebilla de despliegue en un agujero de ajuste.
20	Loop	manual_de_usuario	Lazo	Lazo
21	Use the loop to tuck a	manual_de_usuario	Use el lazo p	Utilizar el lazo para guardar el exceso de la correa de su ZenWatch de ASUS.
22	Strap	manual_de_usuario	Correa	Correa
23	The strap allows you t	manual_de_usuario	La correa le p	La correa le permite llevar su ZenWatch de ASUS en su muñeca.

Figura 6: hoja de datos de la memoria de traducción para la prueba de MT Ranking

En este archivo figuraban 525 segmentos. Una vez preparado, se crea el proyecto en DQF. Para ello, se añade el CSV preparado en la herramienta.

Finalmente, los traductores reciben un correo electrónico generado automáticamente desde la herramienta con un enlace que les permite acceder a la plataforma y comenzar la tarea. A los traductores, al margen de este correo electrónico, se les envió un correo en el que se comentaban las características del proyecto. Se les informaba de en qué consistía la tarea y cómo debían proceder al escoger entre segmentos. La instrucción más relevante que debían tener en cuenta era que debían escoger el segmento que considerasen que, a la hora de poseditar, tendría un menor esfuerzo de posedición (es decir, que tuviese el menor número de ediciones posibles). Asimismo, aunque la herramienta permite pausar la tarea en cualquier momento, se les recomendó que no lo hicieran, para tratar de evitar que se alterase el tiempo de duración. En caso de que los dos segmentos fueran iguales, se les recomendó que escogieran uno al azar, dado que estos no se tendrían en cuenta a la hora de valorar el estudio. Los traductores contaban con aproximadamente una semana para llevar a cabo el ejercicio.

3.5 Prueba 2: prueba de productividad

La prueba de productividad conformó la segunda fase de este proyecto. El objeto de estudio en este caso es la productividad y, una vez más, el método empleado es empírico-experimental. Como se ha visto con anterioridad, esta prueba consiste en que un traductor posedita un texto directamente en la herramienta de DQF. La herramienta le presenta al traductor tanto el segmento de origen como el segmento de destino con la traducción automática en bruto, en la que el traductor debe hacer las modificaciones que considere oportunas según el objetivo que se busca (en este caso, un texto final con una calidad igual a una traducción humana). Así, la segunda fase de este proyecto consiste en que seis de los traductores que llevaron a cabo la primera prueba de MT ranking realicen una prueba de productividad con los mismos textos. A raíz de esta prueba, la herramienta permite determinar el tiempo empleado (en milisegundos) en cada segmento, así como el número de ediciones.

Los resultados que se obtienen se analizan para comprobar tanto la distancia como el esfuerzo de posedición a la hora de trabajar con los dos motores:

> The results provide insight into the difference in time and effort between light and full post-editing. Users will also learn about the impact of certain errors on translation quality, the variance across languages and content types, the correlation with certain metrics and scores or the influence of the translator's profile (age, gender, experience, etc.) on post-editing.

Al igual que en la tarea anterior, a los participantes se les recomendó que no pausaran la prueba para evitar alterar el flujo de trabajo en la medida de lo posible. Los traductores recibieron dos pruebas independientes de posedición. Así, fue necesario crear dos proyectos distintos en la herramienta de DQF. Para evitar que se cometiesen errores o que se *saturasen* al trabajar con el mismo texto o motor, las pruebas se dispusieron de tal forma que ningún traductor poseditase el mismo tipo de texto o el mismo motor las dos veces. El reparto de los textos y motores se hizo de forma totalmente aleatoria y, durante el tiempo en el que los traductores realizaron las pruebas, no se intervino de ninguna manera en el experimento. En la siguiente tabla se puede ver la distribución de las pruebas con los participantes:

Participante	Prueba 1	Prueba 2
Poseditor 1 (P1)	Documentación TAE	Marketing TAN
Poseditor 2 (P2)	Documentación TAE	Marketing TAN
Poseditor 3 (P3)	Documentación TAE	Marketing TAN
Poseditor 4 (P4)	Documentación TAN	Marketing TAE
Poseditor 5 (P5)	Documentación TAN	Marketing TAE
Poseditor 6 (P6)	Documentación TAN	Marketing TAE

Tabla 1: desglose de participantes para las pruebas de productividad

Tras tener la distribución de las tareas realizada, se creó el proyecto en la plataforma de DQF. Para ello, se emplearon las mismas memorias de traducción que se utilizaron a la hora de crear el archivo CSV para la tarea de MT Ranking anterior. Para realizar esta prueba, los traductores tuvieron una semana. Las directrices se presentaron

a través de un correo electrónico, en el que se les explicaba que debían poseditar con el estándar de *equal to human quality* (esto es, *full postediting*). Al igual que en la prueba de MT Ranking, los traductores recibieron un correo para cada prueba con el enlace a DQF. Allí, al traductor se le presentaba cada segmento original (así como el segmento anterior y posterior) con la traducción automática de la herramienta que debían poseditar. Una de las mayores desventajas de esta prueba radicaba en que, a diferencia de en la prueba de evaluación que se verá a continuación, el poseditor no podía modificar el segmento anterior.

3.5.1 Evaluación de calidad

Las pruebas de evaluación de calidad sirvieron como un instrumento más que complementaba las pruebas de productividad, que se realizaron de forma simultánea a la prueba de MT Ranking. Así, se llevaron a cabo cuatro pruebas que consistían en determinar la fluidez y la precisión de los mismos segmentos con los que los traductores estaban trabajando en ese mismo momento. En este caso, las cuatro pruebas que correspondían a la determinación de la calidad las realicé yo misma. Aunque lo ideal hubiese sido que al menos dos traductores ajenos al resto de pruebas las hubiesen realizado, debido a la falta de recursos, esto no fue posible. De esta manera, el evaluador de las pruebas fue un traductor en el mismo rango de edad y con una formación similar (graduado en Traducción e Interpretación, máster en Tradumática y experiencia de varios años en el sector) que el grupo de seis y nueve utilizados con anterioridad, pero que no había realizado la prueba de MT Ranking y que tampoco trabajaría directamente en las pruebas de productividad.

Para realizar esta evaluación de calidad, se crearon cuatro pruebas distintas en DQF, una para cada posible combinación: TAE – Marketing, TAN – Marketing, TAE – Documentación y TAN – Documentación.

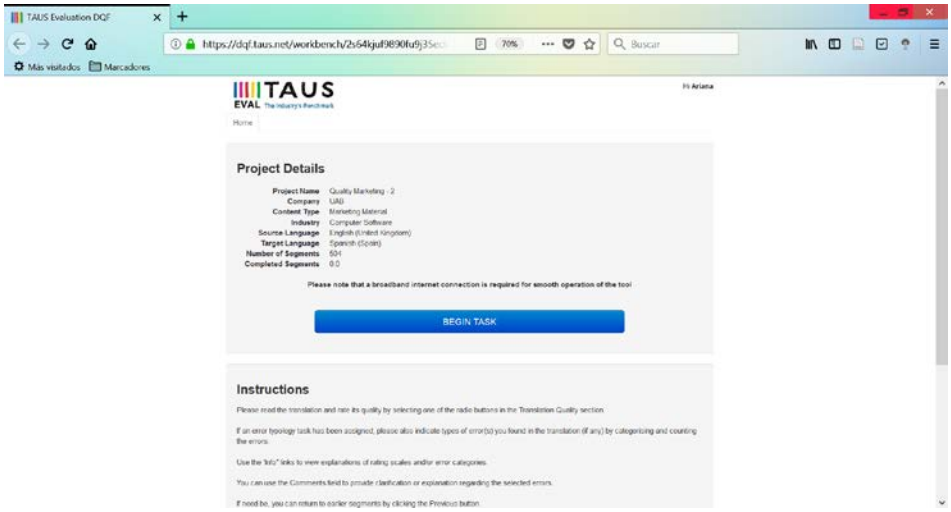


Figura 7: pantalla de inicio de una de las pruebas de calidad en DQF

En estas cuatro pruebas, el traductor debía evaluar tanto la fluidez como la precisión simultáneamente en cuatro grados distintos, dependiendo del criterio que se estuviera evaluando. En términos de fluidez, se evalúa desde Incomprensible (1) a Sin errores (4), pasando por Poco fluido (2) y Bueno (3):

Flawless	*refers to a perfectly flowing text with no errors.*
Good	*refers to a smoothly flowing text even when a number of minor errors are present.*
Disfluent	*refers to a text that is poorly written and difficult to understand.*
Incomprehensible	*refers to a very poorly written text that is impossible to understand.*

Tabla 2: descripción de la escala de grados de fluidez de DQF

En cuanto a la precisión, la escala abarca también cuatro grados: Todo (4), Bastante (3), Poca (2) y Nada (1). El evaluador pasaba por cada segmento y le proporcionaba una puntuación según la calidad que presentase.

Everything	*All the meaning in the source is contained in the translation, no more, no less.*
Most	*Almost all the meaning in the source is contained in the translation.*
Little	*Fragments of the meaning in the source are contained in the translation.*
None	*None of the meaning in the source is contained in the translation.*

Tabla 3: descripción de la escala de grados de precisión de DQF

Una vez finalizada la prueba, se obtiene la siguiente pantalla, en la que se avisa que no se podrán editar los resultados:

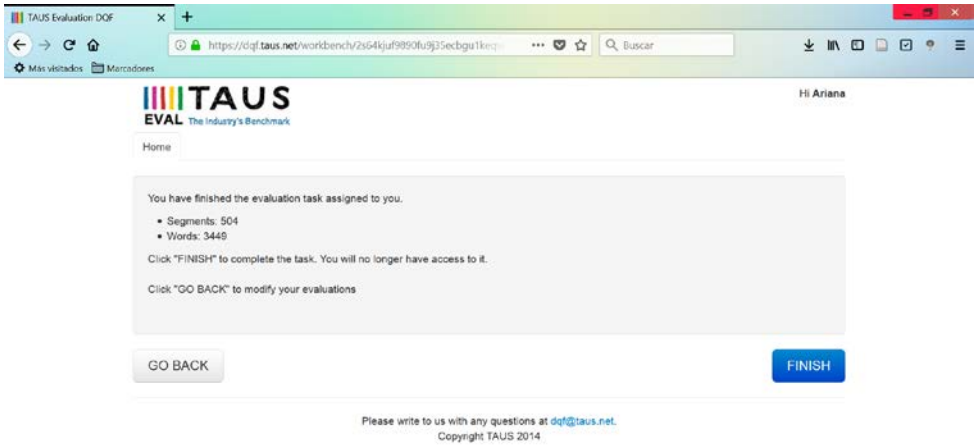


Figura 8: aviso de TAUS de finalización de una prueba

Mediante las pruebas de evaluación de la calidad se obtiene un criterio cualitativo (aunque los resultados de esta prueba sean cuantitativos), que en este caso consiste en la percepción de un único traductor sobre la calidad. Así, este criterio se puede cotejar en el apartado de resultados, como se verá más adelante, con el criterio cuantitativo, esto es, el número de palabras, para obtener un mayor abanico de resultados.

4 Resultados obtenidos

Los resultados que se consiguieron tras realizar las pruebas anteriormente mencionadas fueron varios archivos HTML, que se corresponden con cada uno de los participantes de las pruebas de DQF que se descargaron directamente de la plataforma, así como los seis archivos CSV de las pruebas de productividad. Con respecto a la prueba de MT Ranking, también se generó un archivo HTML en el que figuraban las estadísticas generales de todas las pruebas. Con respecto a la estructura de este capítulo, se divide en dos grandes apartados: una primera parte que consiste en el análisis de los resultados y una segunda, que trata la interpretación de los resultados obtenidos. Sobre estos últimos, se desglosan de la siguiente manera: para la prueba de MT Ranking se genera un archivo HTML descargable con todos los resultados de los participantes, así como un archivo CSV con sus opciones para cada uno de ellos. En cuanto a las pruebas de productividad, se generaron doce archivos CSV en los que se especifican el número de ediciones, el tiempo empleado y el resultado final al que llegaron los participantes (los segmentos de destino poseditados). Con relación a las pruebas de calidad, como un instrumento más, se generaron cuatro archivos CSV en los que se especifican la puntuación otorgada a cada uno de los segmentos de traducción automática en bruto.

4.1 Análisis de los resultados

El procedimiento de análisis de los resultados empleado consistió, en primer lugar, en descargar toda la información pertinente en distintas hojas de CSV de la plataforma de DQF, así como los archivos HTML con la información relevante de las acciones tomadas por cada traductor participante. Tras la descarga, se crearon dos nuevas hojas de datos. En primer lugar, se construyó una única hoja de datos en la que se volcó toda la información obtenida a partir de las cuatro pruebas de calidad, a saber: calidad de TAE del texto de marketing, calidad de TAE del texto de documentación, calidad de TAN del texto de marketing y calidad de TAN del texto de documentación. En ellas aparecía tanto el segmento original y el segmento de destino para cada motor y la puntuación otorgada a cada uno de los segmentos para fluidez y precisión. Esta hoja de datos estará disponible en los anexos.

Asimismo, se creó un segundo archivo de hoja de datos en la que se volcaron todas las columnas descargadas de las hojas originales de cada poseditor en cada una de las pruebas de productividad. En este caso, se prepararon dos hojas en el mismo archivo, una para los resultados de los textos de marketing y otros para los de documentación del usuario. En las dos hojas se siguió la misma estructura: en primer lugar, el segmento original en inglés, seguido por el número de palabras. Tras esto, el segmento de destino del motor de TAE, a saber, la opción que proporciona Microsoft Translator. Las siguientes columnas representan la opción poseditada del traductor, el número de palabras en el segmento de destino, el tiempo de posedición (en milisegundos) y la distancia de edición[4]. Una vez terminada esta estructura, se añadió a la derecha los resultados del segmento de destino del motor de TAN (esto es, la opción que proporciona Google Translate) y se continuó la misma estructura empleada que con el sistema de TAE: la opción poseditada del traductor, el número de palabras en el segmento de destino, el tiempo de posedición (en milisegundos) y la distancia de edición.

Para realizar un mejor procesamiento de los datos y obtener unos mejores resultados, que abarcasen simultáneamente los de calidad obtenidos en las pruebas de calidad realizadas y los de las pruebas de productividad, se crearon tres categorías distintas. Tal y como se explicaba anteriormente al describir las pruebas de evaluación de calidad, con sus distintos niveles, DQF otorga una puntuación de 1 a 4, siendo uno la puntación más baja y cuatro, la más alta, tanto para la evaluación de fluidez como de precisión. Partiendo de esta base, se interrelacionaron todos los datos obtenidos y se establecieron distintas categorías, según su calidad. A continuación, se muestran las tablas con los resultados interrelacionados entre fluidez y precisión.

La primera categoría, que se ha denominado «categoría de mayor calidad», se centraba en los segmentos de calidad alta, lo que equivaldría a obtener alguno de los siguientes resultados:

[4] La herramienta de TAUS denomina la distancia de edición como «editing effort», esfuerzo de edición. Dado que el esfuerzo implica una actividad cognitiva, que no se mide en este proyecto, aquí se habla en todo momento de «distancia de edición», puesto que es más preciso.

Fluidez[5]	Precisión[6]
Sin errores	Todo
Sin errores	Bastante
Bueno	Todo
Bueno	Bastante

Tabla 4: descripción de resultados posibles para la categoría de mayor calidad

La segunda categoría, denominada «categoría de calidad intermedia», son los segmentos de calidad media, que equivaldría a obtener alguno de los siguientes resultados:

Fluidez	Precisión
Sin errores	Nada
Incomprensible	Todo
Sin errores	Poco
Poco fluido	Todo
Poco fluido	Bastante
Bueno	Poco
Incomprensible	Bastante
Bueno	Nada

[5] Es necesario recordar que, en fluidez, 4 equivale a Sin errores, 3 a Buena, 2 a Poco fluido y 1 a Incomprensible.
[6] En el caso de la precisión, los grados son Toda (4), Bastante (3), Poco (2) y Nada (1).

Tabla 5: descripción de resultados posibles para la categoría de calidad intermedia

En último lugar se encuentra la tercera categoría, denominada «categoría de calidad baja», en la que figuran los segmentos de muy mala calidad y que ya no merecería la pena poseditar porque el esfuerzo sería demasiado alto. En este caso, sería necesario que apareciera alguno de estos resultados:

Fluidez	Precisión
Poco fluido	Nada
Incomprensible	Poco
Poco fluido	Poco
Incomprensible	Nada

Tabla 6: descripción de resultados posibles para la categoría de calidad baja

Al mismo tiempo, se clasificaron los segmentos según su longitud en el segmento original: una longitud menor a cinco palabras (segmentos cortos), una longitud de seis a diecinueve palabras (segmentos medios) y más de veinte palabras (segmentos largos).

El primer análisis de los datos consistió en extraer los tiempos de posedición y el número de ediciones de media de cada participante. Para ello, se empleó la fórmula CONTAR.SI, de tal forma que se calculase el tiempo de posedición y el número de ediciones según las categorías mencionadas anteriormente: por un lado, la calidad de los segmentos de traducción automática en bruto (alta, media o baja), así como la longitud (menos de cinco palabras, de seis a diecinueve palabras y más de veinte palabras). Estos datos se obtuvieron únicamente para conocer de forma general la actividad de cada traductor en las distintas pruebas. No se desarrollarán en los siguientes análisis.

Para obtener los resultados generales, independientemente de los participantes, se creó una hoja de datos distinta en la que se categorizaron todos los datos

simultáneamente. En la primera hoja se encontraba, en la primera columna, el segmento en origen en inglés y, tras esto, el número de palabras y el tipo de motor que se corresponde con el segmento de origen. En las columnas D y E figuraba el valor otorgado en la evaluación de calidad. En las columnas E y F finalmente el tiempo de posedición y la distancia de edición, respectivamente.

4.2 Interpretación de los resultados

Para este apartado, se interpretarán los resultados siguiendo el mismo orden cronológico en el que se efectuaron las pruebas. En primer lugar, se tratarán los resultados obtenidos en el MT Ranking. Tras esto, se pasará a los datos de las pruebas de productividad.

4.2.1 MT Ranking

Al analizar los resultados de la prueba de MT Ranking, se puede observar que los participantes escogieron en 7 de cada 10 casos la traducción automática neuronal sobre la estadística. Es necesario recordar que la pregunta que se les hacía en esta prueba era qué opción de las dos presentadas consideraban que tendría menos ediciones a la hora de poseditar. Así, en la mayor parte de los casos, creían que la opción con un menor número de ediciones era la presentada por el motor neuronal de Google. Esto se ve en el gráfico que se presenta a continuación:

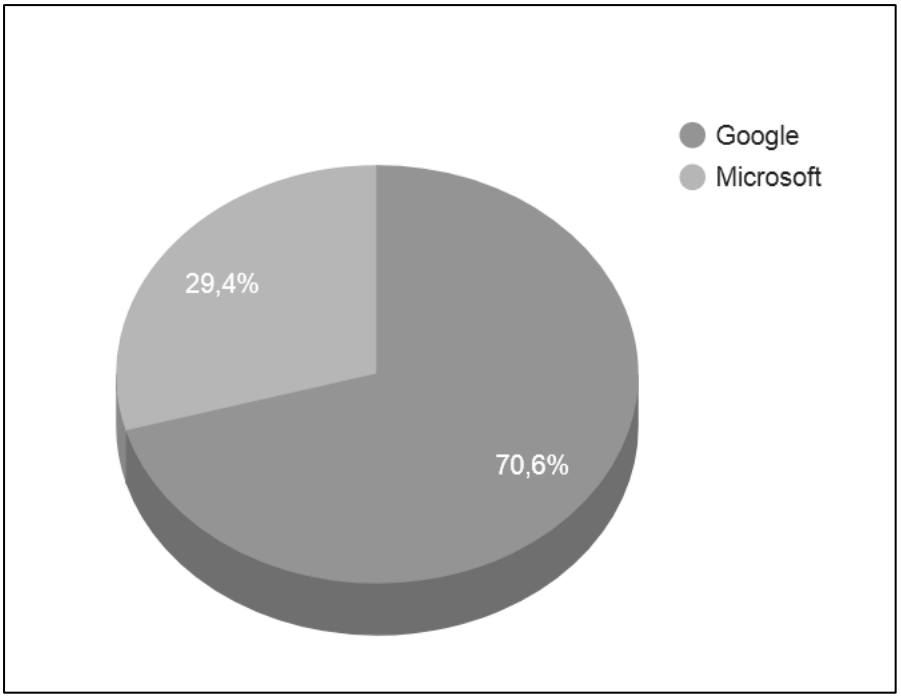


Figura 9: gráfica de la prueba de MT Ranking

4.2.2 Prueba de evaluación de calidad

Tal y como se comentaba en el apartado de análisis de los resultados, para la prueba de productividad se efectuaron dos tipos de análisis distintos. En primer lugar, se trabajó con los resultados obtenidos a través de la evaluación de calidad, esto es, tanto de la fluidez como de la precisión. A pesar de que esta se considera un instrumento más para la obtención de los datos de productividad, los datos que aportan esta evaluación son relevantes.

Para calcularlos, se prepararon las hojas de datos a partir de los resultados obtenidos y se introdujeron en el programa R, mediante el cual se calcularon las medias correspondientes. A continuación, se presentan los resultados en forma de gráficas. En ellas se puede cotejar las distintas categorías y puntuaciones otorgadas a los segmentos de la traducción automática en la prueba de evaluación, de forma porcentual sobre el total (100 %). Por tanto, lo que se observa es el porcentaje que le corresponde a un motor o a otro para las distintas categorías.

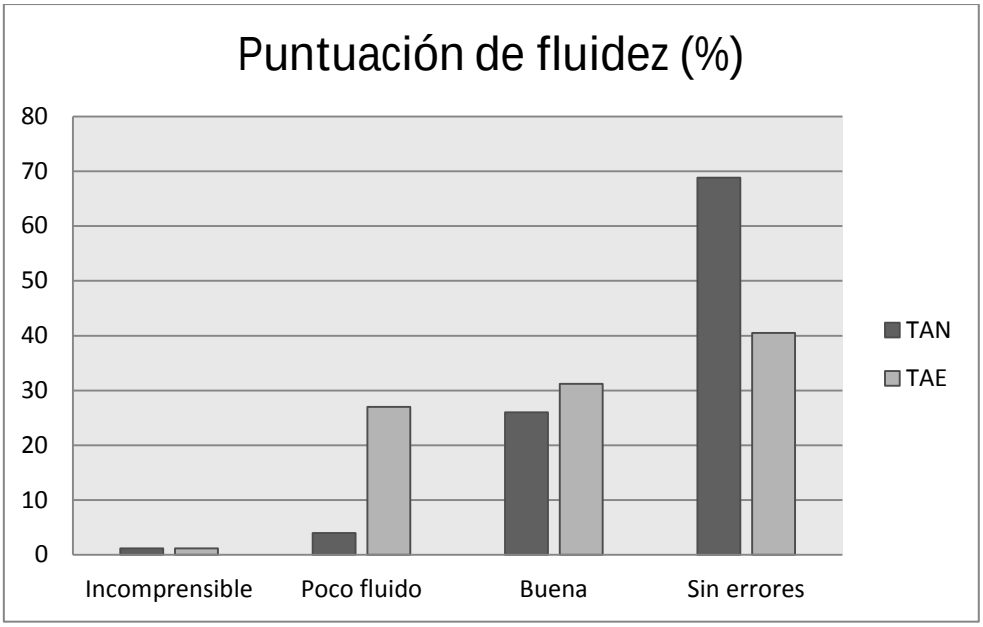


Figura 10: gráfica de puntuación de la prueba de evaluación de fluidez (en %)

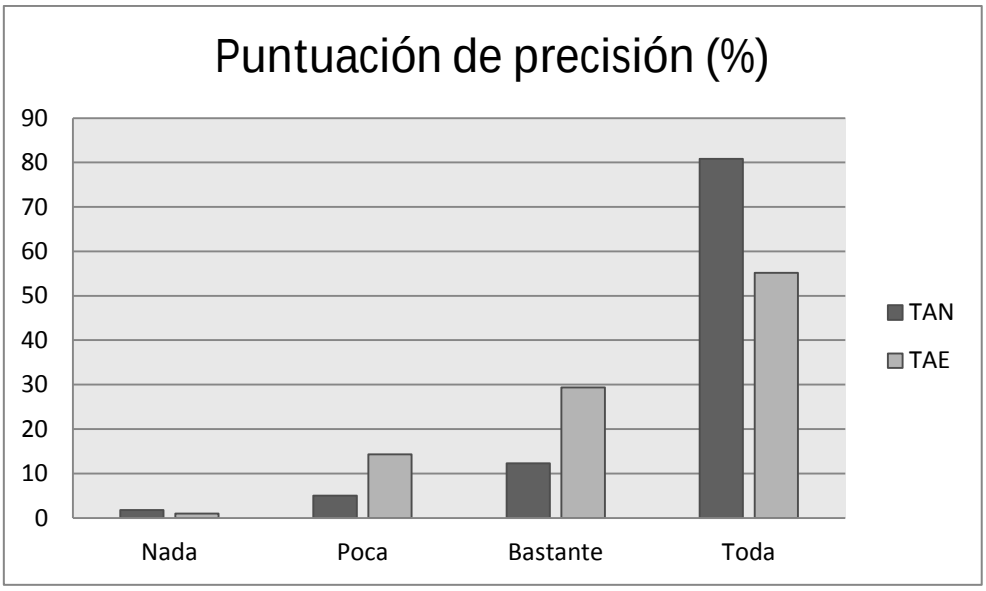


Figura 11: gráfica de puntuación de la prueba de evaluación de precisión (en %)

Tal y como se puede observar en la gráfica 10, el motor de TAN obtiene mejores resultados en términos de fluidez que el motor de TAE, ya que en la puntuación «Sin errores» obtiene unos valores superiores. Sucede exactamente lo mismo en la gráfica de precisión, en la que la puntuación «Toda» también es superior. Con lo cual, el motor de

TAN obtiene mejores resultados tanto en fluidez como en precisión que el de TAE cuando la valoración de la calidad es la más alta.

No obstante, esto no ocurre en el segundo nivel, que se corresponde con una fluidez «Buena» y «Bastante» precisión, tal y como se puede observar en las dos gráficas. En este nivel, el motor de TAE es considerablemente superior que el de TAN. Sucede lo mismo en el tercer nivel: el motor de TAE es, de nuevo, superior al de TAN. Esto queda especialmente patente en la gráfica de fluidez, donde el motor de TAE tiene un porcentaje que roza el 30 % de los resultados, mientras que el de TAN apenas llega al 5 %.

En lo que se refiere a la categoría más baja, ambos motores presentan unos resultados muy igualados en temas de calidad. Sin embargo, es importante ver que, aunque en el caso de la fluidez los valores son exactamente iguales, en la precisión el motor de TAN tiene un valor porcentual ligeramente superior que el de TAE.

4.2.3 Prueba de productividad

Como se comentaba en el apartado de análisis de los resultados, para la prueba de productividad se preparó un archivo de hoja de datos con todos los datos volcados según la categoría. Tras esto, se usó la herramienta R para extraer los resultados estadísticos de cada una de las categorías. Mediante este programa se calcularon tanto las medias como las desviaciones de los resultados totales obtenidos en cada categoría, tanto para el tiempo de posedición como para la distancia de edición. Extrajimos una primera descripción estadística que consistía en la media, mediana, desviación estándar y la distribución en cuartiles. A continuación, llevamos a cabo un análisis comparativo entre pares de datos independientes. Para ello, primero determinamos si la variancia era probabilística mediante un test de Levene. A continuación, calculamos la interrelación entre los pares de datos mediante una t de Student. Estos resultados se presentan a continuación tanto en forma de diagrama de cajas[7] como en tablas que los acompañan.

[7] En el caso de los gráficos presentados en este apartado, todos los datos figuran en inglés, ya que es así como figuraban en la herramienta de DQF. Es por este motivo que, en este caso, figura edit effort, en lugar de edit distance.

En primer lugar, se abordarán los resultados de la distancia de edición, para después continuar con los del tiempo de posedición. En este caso, a pesar de que en las gráficas figura el tiempo en milisegundos, a lo largo de este apartado se hablará en segundos, para facilitar la comprensión. En los subapartados siguientes se procede a la interpretación de los datos.

4.2.3.1 Categoría de mayor calidad

A continuación, se muestran las gráficas y la interpretación de los resultados para la categoría de mayor calidad:

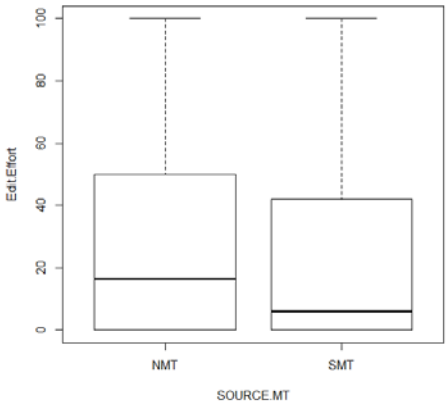


Figura 12: distancia de edición de la categoría de mayor calidad, menos de cinco palabras

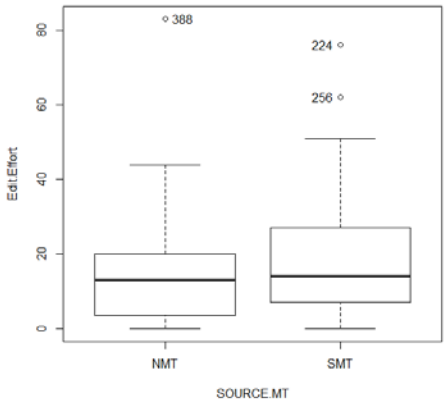


Figura 13: distancia de edición de la categoría de mayor calidad, de seis a diecinueve palabras

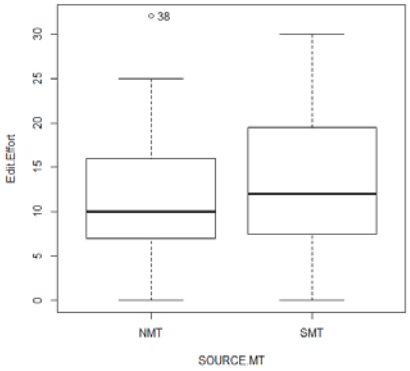


Figura 14: distancia de edición de la categoría de mayor calidad, más de veinte palabras

		Menos de cinco palabras		De seis a diecinueve palabras		Más de veinte palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Distancia de edición	TAN	28,23874	30,59252	13,63406	11,21979	11,59259	7,933664
	TAE	21,88288	27,73378	17,65217	13,39398	13,59259	7,855715
Test de Levene (p-valor)		> 0,05		< 0,05		> 0,05	
T de Student (p-valor)		0,02229		0,0001488		0,3563	

Tabla 7: resumen de los datos estadísticos obtenidos para la categoría de mayor calidad para la distancia de edición

La figura 12 representa la distancia de edición de la traducción automática neuronal (TAN, en adelante) y la traducción automática estadística (TAE, en adelante) en la categoría uno (alta calidad) con segmentos que tienen menos de cinco palabras. Por su parte, las figuras 13 y 14 representan los mismos datos, pero en los segmentos de seis a diecinueve palabras y más de veinte palabras, respectivamente.

Como se puede ver en la gráfica de la figura 12, la distancia de edición es superior en el caso de la TAN, mientras que el de la TAE es considerablemente inferior. Esto se ve reflejado en las medias (28,23 para el motor de TAN y 21,88 para el motor de TAE). En cuanto al test de Levene, el p-valor es superior a 0,05, lo que determina que la distribución de los datos es probabilística. A continuación, aplicamos una t de Student para datos probabilísticos que arrojó un p-valor inferior a 0’05 (0,02229). Por lo tanto, podemos determinar que la distancia de edición es significativamente mayor al usar el motor de TAN que de TAE.

En el caso de la figura 13, que muestra los valores de distancia de edición de la TAN y la TAE en los segmentos de alta calidad de seis a diecinueve palabras, se puede observar a simple vista que hay tres valores atípicos. La media para la distancia de edición de la TAN es de 13,63, mientras que presenta un valor de 17,65 para el de TAE. En cuanto al test de Levene, el p-valor, a diferencia del resultado obtenido en el caso anterior, es menor que 0,05, lo que implica que la distribución de los datos es probabilística. La prueba de t de Student para datos probabilísticos arroja un valor considerablemente bajo (0,0001488). Así, a diferencia del caso anterior, aquí se puede determinar que la distancia de edición es significativamente mayor cuando se usa TAE que cuando se emplea TAN.

En lo que se refiere a la figura 14, que muestra los valores de distancia de edición de la TAE y la TAN en los segmentos de calidad alta que tienen más de 20 palabras, en este caso solo hay un valor atípico que se ve en el uso del motor de TAN. La media de distancia de edición para este motor es de 11,59, mientras que para el motor de TAE es de 13,59. El p-valor que se obtiene con el test de Levene es mayor que 0,05, con lo cual la distribución de los datos es probabilística. El p-valor que se obtiene al calcular la prueba de t de Student es de 0,3563. En este caso, la distancia de edición es superior al usar un motor de TAE que uno de TAN, pero, a diferencia de los dos casos anteriores, esta no es una diferencia estadísticamente significativa.

A continuación, se muestran las gráficas obtenidas para las mismas categorías (segmentos de alta calidad para los segmentos de menos de cinco palabras, de seis a diecinueve y más de veinte), pero calculados con respecto al tiempo de posedición:

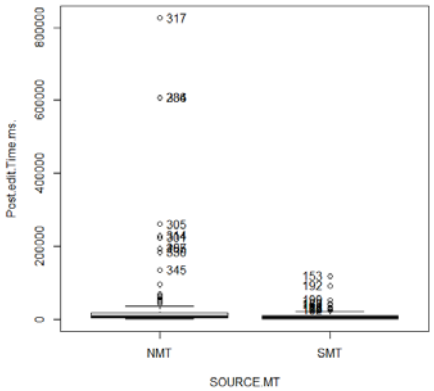


Figura 15: tiempo de posedición (en ms) de la categoría de mayor calidad, menos de cinco palabras

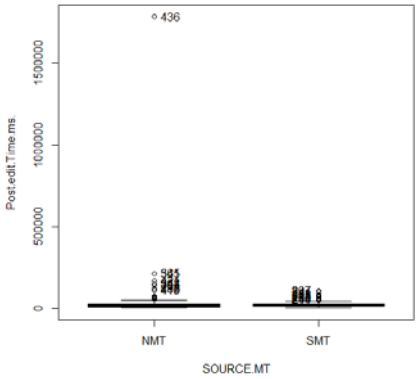


Figura 16: tiempo de posedición (en ms) de la categoría de mayor calidad, de seis a diecinueve palabras

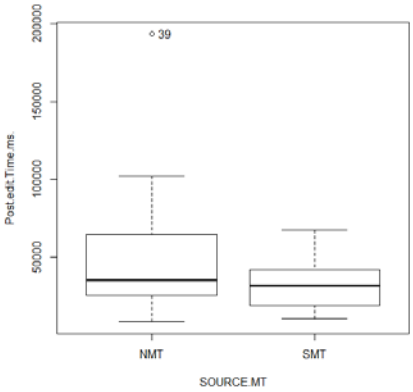


Figura 17: tiempo de posedición (en ms) de la categoría de mayor calidad, más de veinte palabras

		Menos de cinco palabras		De seis a diecinueve palabras		Más de veinte palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Tiempo de posedición (ms)	TAN	26855,32	85702,81	29600,65	109757,01	47355,19	38835,29
	TAE	8395,45	12208,37	19638,44	14704,32	32549,26	15603,02
Test de Levene (p-valor)		< 0,05		> 0,05		> 0,05	
T de Student (p-valor)		0,001691		0,1356		0,07175	

Tabla 8: resumen de los datos estadísticos obtenidos para la categoría de mayor calidad para el tiempo de posedición

A diferencia de las gráficas representadas para la distancia de edición de la categoría anterior, a simple vista se puede observar en estas que se dan muchos más valores atípicos (algunos de ellos incluso se solapan). Esto sucede especialmente en las figuras 15 y 16.

La figura 15, donde los segmentos tienen menos de cinco palabras, presenta una gran cantidad de valores atípicos, especialmente concentrados en los cuartiles primero y segundo. La media en este caso es de 26,85 segundos en el caso de la TAN, más del triple que para los resultados del motor de TAE, de 8,35 segundos. El test de Levene resultó en un p-valor inferior a 0,05 (la distribución de datos no es probabilística), mientras que el test de Student arrojó un resultado de 0,001691. Así, el tiempo de posedición para los segmentos de alta calidad que tienen menos de cinco palabras es significativamente más alto para la TAN que para la TAE.

En la figura 16, al igual que en la 15, los valores atípicos se concentran en el primer cuartil, incluso más que en la figura 15. El tiempo de posedición de media en este caso es de 29,60 segundos para el caso de la TAN, mientras que es de 19,63 segundos para la TAE. Al contrario que en los resultados obtenidos para los segmentos de menos de cinco palabras, el test de Levene dio un valor superior a 0,05, con lo cual la distribución de datos es probabilística. La prueba de Student resultó en un valor de 0,1356. Aunque el tiempo de posedición con TAN es mayor que el de TAE, no se puede afirmar que la diferencia sea estadísticamente significativa.

En el caso de la figura 17, se muestran los resultados obtenidos para los segmentos de calidad alta con un número de palabras superior a veinte. En este caso, los valores son mucho más homogéneos. El tiempo de posedición de media para los resultados con TAN es de 47,35 segundos, mientras que el de TAE es de 32,54 segundos. El test de Levene dio un resultado superior a 0,05 (la distribución de los datos es probabilística), mientras que el de Student mostró un valor de 0,07175. Así, el tiempo de posedición con TAN también es mayor que el de TAE, pero una vez más no se puede decir que esta diferencia sea estadísticamente significativa. Sería necesario recabar una mayor cantidad de datos.

Para facilitar la comprensión global de todos los resultados, se expresan a continuación en forma de tabla:

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

Tabla 9: resumen de los datos recogidos de la categoría de mayor calidad

En resumen, es necesario destacar que la distancia de edición es superior para la TAN cuando los segmentos tienen menos de cinco palabras, pero es mayor en el caso de la TAE para los segmentos de seis a diecinueve palabras. Sería necesario recabar más resultados para saber qué sucede con los segmentos de más de veinte palabras. En el caso del tiempo de posedición, este es mayor en los tres casos para la TAN, pero en ningún caso estos resultados son significativos, ya que las muestras son demasiado heterogéneas. Sería necesario recopilar más datos y volver a efectuar las pruebas.

4.2.3.2 Categoría de calidad intermedia

En este apartado se analizan los resultados de la categoría dos, esto es, los segmentos que obtuvieron una calidad media en la prueba de evaluación. Se seguirá el mismo orden que en el apartado anterior:

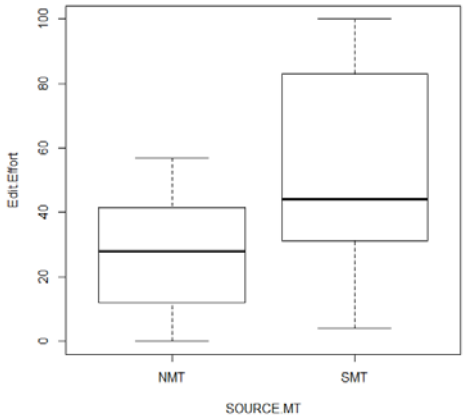


Figura 18: distancia de edición de la categoría de calidad intermedia, menos de cinco palabras

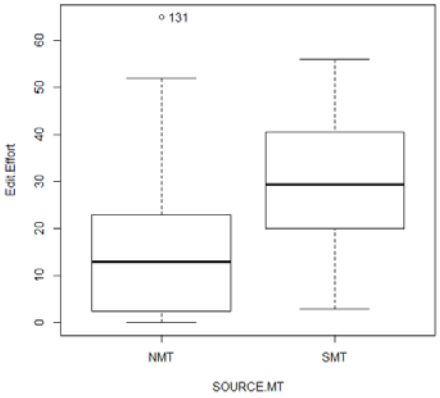


Figura 19: distancia de edición de la categoría de calidad intermedia, de seis a diecinueve palabras

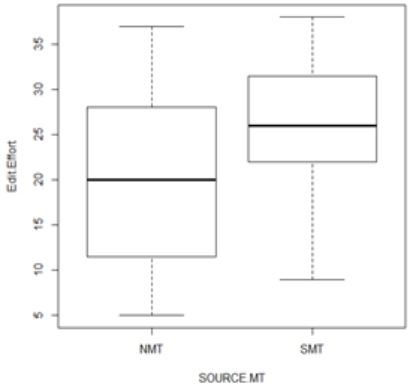


Figura 20: distancia de edición de la categoría de calidad intermedia, más de veinte palabras

		Menos de cinco palabras		De seis a diecinueve palabras		Más de veinte palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Distancia de edición	TAN	27,16667	20,09447	15,35714	13,62549	19,75000	9,789837
	TAE	52,91667	32,70101	29,96429	12,24100	26,33333	8,150107
Test de Levene (p-valor)		> 0,05		> 0,05		> 0,05	
T de Student (p-valor)		0,03184		1,079e-11		0,08718	

Tabla 10: resumen de los datos estadísticos obtenidos para la categoría de calidad intermedia para la distancia de edición

En la figura 18 se ven los resultados de la distancia de edición en los segmentos de calidad media que tienen menos de cinco palabras. La media de la TAN es inferior a la de la TAE (27,16 y 52,91, respectivamente). El p-valor del test de Levene es superior a 0,05 (es una distribución probabilística). A continuación, se realizó una prueba de Student, mediante la que se obtuvo un valor de 0,03184. Así, la distancia de edición es significativamente menor usando el motor de TAN que de TAE, al igual que sucedía en los segmentos de la categoría de mayor calidad.

En la figura 19 se puede observar la gráfica de los resultados obtenidos para la distancia de edición de los segmentos de seis a diecinueve palabras. En este caso, la distancia de edición es superior en la TAE que en la TAN, con un valor 29,96 para la primera y uno de 15,35 para la segunda. Por otra parte, el test de Levene da un valor superior a 0,05 (una distribución de los datos probabilística). Tras esto, se llevó a cabo una t Student, que arrojó un valor de 1,079e-11. La distancia es entonces significativamente menor al usar el motor de TAN que el de TAE, al contrario de lo que sucede en los segmentos con calidad alta. En la otra ocasión en la que sucede esto (los segmentos de más de veinte palabras de alta calidad), la muestra de datos era demasiado heterogénea para confirmarlo.

Por lo que respecta a la figura 20, en ella se muestra la gráfica con los resultados de calidad media con más de veinte palabras. La media es 19,75 ediciones para los segmentos de TAN y de 26,33 para los de TAE. El p-valor del test de Levene es superior a 0,05, lo que implica que la distribución de datos es probabilística. A continuación, se realizó el test de Student, que arrojó un resultado de 0,08718. La distancia de edición es inferior al trabajar con el motor de TAN que con el de TAE, pero esta diferencia no es estadísticamente significativa.

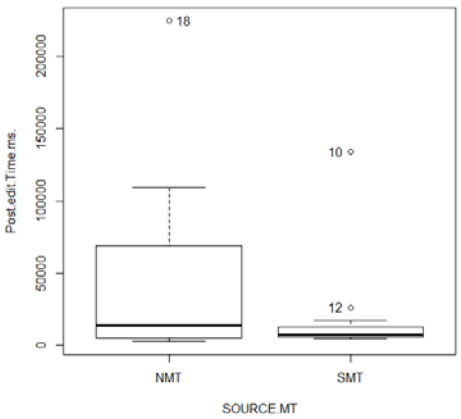


Figura 21: tiempo de posedición (en ms) de la categoría de calidad intermedia, menos de cinco palabras

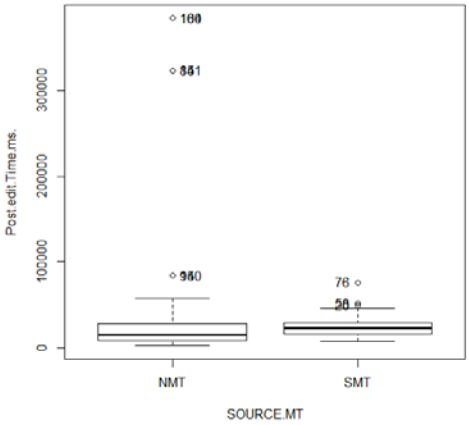


Figura 22: tiempo de posedición (en ms) de la categoría de calidad intermedia, de seis a diecinueve palabras

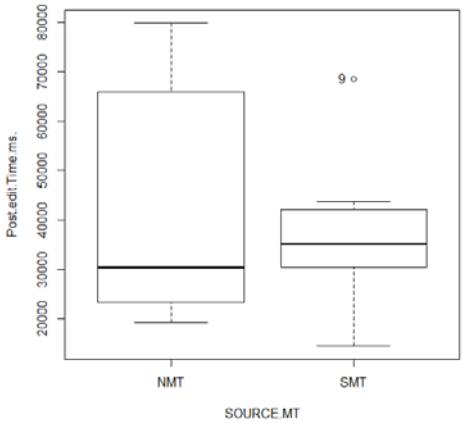


Figura 23: tiempo de posedición (en ms) de la categoría de calidad intermedia, más de veinte palabras

		Menos de cinco palabras		De seis a diecinueve palabras		Más de veinte palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Tiempo de posedición	TAN	45193,33	68355,15	35932,62	73592,30	42495,83	23961,08
	TAE	19656,67	36493,29	23356,07	11284,35	36512,50	12974,72
Test de Levene (p-valor)		> 0,05		< 0,05		> 0,05	
T de Student (p-valor)		0,2696		0,1252		0,4549	

Tabla 11: resumen de los datos estadísticos obtenidos para la categoría de calidad intermedia para el tiempo de posedición

La figura 21 muestra la gráfica de resultados que se corresponden con los segmentos de la categoría dos (calidad media) con un número de palabras igual o inferior a cinco. Al igual que sucedía con las gráficas de tiempo de posedición de la categoría uno (calidad alta), aparecen muchos valores atípicos, especialmente concentrados en el primer cuartil. En lo que respecta a las medias de tiempo, la media de TAN es considerablemente más alta que la de TAE (45,19 frente a 19,65). El p-valor del test de Levene es superior a 0,05, mientras que el p-valor del test de Student es de 0,2696. Así, el tiempo de posedición es mayor al usar el motor de TAN que el de TAE, pero no de una forma estadísticamente significativa.

Por su parte, la figura 22 representa la gráfica de los resultados de segmentos de categoría media con un número de palabras de seis a diecinueve. Se vuelven a ver una serie de valores atípicos, que implican que ambas muestras son muy heterogéneas. La media en este caso es de 35,93 segundos para el motor de TAN y de 23,35 segundos para el motor de TAE. El p-valor del test de Levene es inferior a 0,05, así que la distribución de datos no es probabilística. Tras este, se efectuó la prueba de Student, que arrojó un valor de 0,1252. Al igual que en el caso anterior, el tiempo de posedición es mayor al usar el motor de TAN que el de TAE, pero, de nuevo, no de una forma estadísticamente significativa.

La figura 23, la gráfica muestra valores homogéneos. Se exponen los resultados de la categoría media para segmentos de más de veinte palabras. En este caso, solo hay un valor atípico en la muestra de TAE. La media para los segmentos de TAN es de 42,49 segundos, superior a la de TAE, de 36,51 segundos. Para el test de Levene, el p-valor es superior a 0,05, lo que implica que la distribución de los datos es probabilística. El valor que arrojó la prueba de Student efectuada posteriormente es de 0,4549. De nuevo, el tiempo de posedición es mayor al usar el motor de TAN que el de TAE, pero no de una forma estadísticamente significativa.

De nuevo, para facilitar la comprensión general de todos los datos recogidos, se presentan en forma de tabla a continuación:

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

Tabla 12: resumen de los datos recogidos de la categoría de calidad intermedia

En el caso de la categoría de calidad intermedia, en todas las ocasiones la distancia de edición es inferior al usar el motor de TAN, pero no el de TAE, aunque solo en dos de los tres casos los resultados son estadísticamente significativos. Con respecto al tiempo de posedición, el tiempo de posedición es siempre más alto al emplear el motor de TAN que el de TAE. No obstante, estos resultados no son significativos, dado que las muestras son muy heterogéneas. En el único caso en el que se contrastaron los resultados es en la categoría de calidad alta, presentada al principio, con segmentos de menos de cinco palabras (figura 15).

4.2.3.3 Categoría de calidad baja

A continuación, se presentan los resultados extraídos para los segmentos de calidad baja:

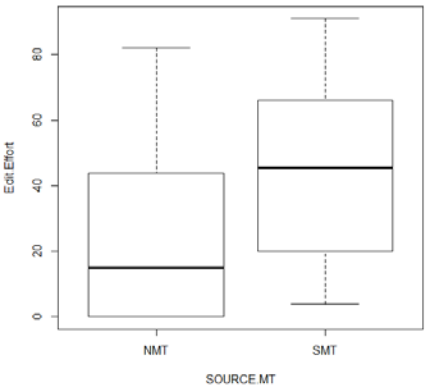


Figura 24: distancia de edición de la categoría de calidad baja, menos de cinco palabras

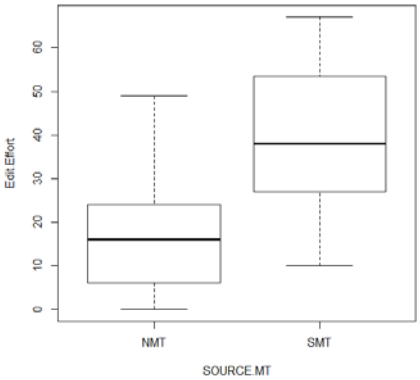


Figura 25: distancia de edición de la categoría de calidad baja, de seis a diecinueve palabras

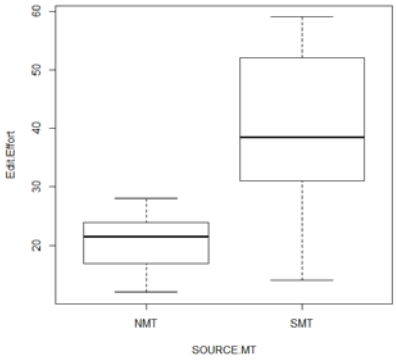


Figura 26: distancia de edición de la categoría de calidad baja, más de veinte palabras

		Menos de cinco palabras		De seis a diecinueve palabras		Más de veinte palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Distancia de edición	TAN	25,70000	29,76129	16,18667	12,43327	20,66667	5,645057
	TAE	44,46667	26,13853	38,60000	15,68525	38,83333	15,942605
Test de Levene (p-valor)		> 0,05		< 0,05		> 0,05	
T de Student (p-valor)		0,01196		< 2,2e-16		0,02511	

Tabla 13: resumen de los datos estadísticos obtenidos para la categoría de calidad baja para la distancia de edición

En la figura 24 se puede ver la gráfica que representa los resultados de la distancia de edición de la categoría tres (calidad baja) de los segmentos de menos de cinco palabras. La distancia de edición de media para el motor de TAN es de 25,70, mientras que es de 44,46 para el motor de TAE. Al realizar el test de Levene, el p-valor dio un resultado mayor que 0,05, lo que implica que la distribución de los datos es probabilística. La t de Student arrojó un valor de 0,01196. Así, queda patente que la distancia de edición de TAN es significativamente inferior que la de TAE.

En lo que se refiere a la figura 25, se muestra la gráfica de los resultados para esta misma categoría con los segmentos de seis a diecinueve palabras. Al igual que para los segmentos de menos de cinco palabras, la distancia es inferior en el caso del motor de TAN que del de TAE (16,18 frente a 38,60). El test de Levene dio un valor inferior a 0,05, mientras que la prueba de Student mostró un valor considerablemente bajo (< 2,2e-16). De nuevo, la distancia de edición es significativamente inferior al usar el motor de TAN que el de TAE.

Con respecto a la figura 26, nos muestra la gráfica para los segmentos de más de veinte palabras. Al igual que en los segmentos de menos de cinco y de seis a diecinueve palabras presentados anteriormente, la distancia de edición es inferior en el caso de la TAN (20,66) que de la TAE (38,83). El p-valor obtenido en el test de Levene, superior a 0,05, implica que la distribución de datos no es probabilística. A continuación, se realizó el test de Student, que reveló un valor de 0,02511. Esto ratifica que la distancia de edición es significativa.

Por último, pasamos a analizar los resultados relacionados con el tiempo de posedición para esta categoría:

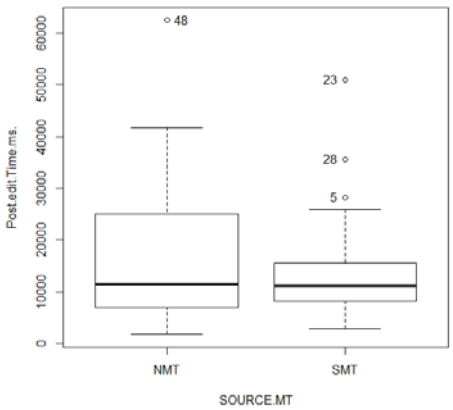


Figura 27: tiempo de posedición (en ms) de la categoría tres, menos de cinco palabras

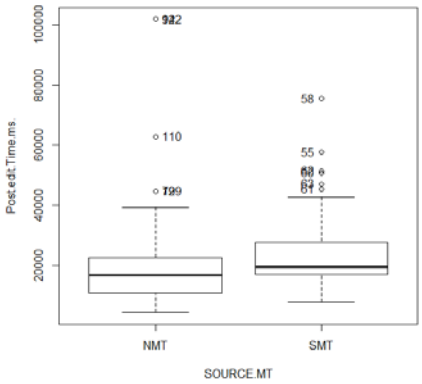


Figura 28: tiempo de posedición (en ms) de la categoría tres, de seis a diecinueve palabras

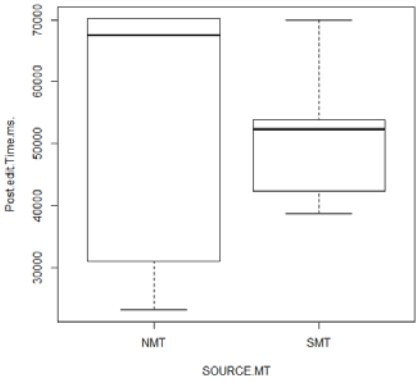


Figura 29: tiempo de posedición (en ms) de la categoría tres, más de veinte palabras

		Menos de cinco palabras		De seis a diecinueve palabras		Más de veinte palabras	
		Media	Desviación	Media	Desviación	Media	Desviación
Tiempo de posedición	TAN	15787,00	12967,66	20528,00	17465,07	54923,33	21741,94
	TAE	14237,67	10148,74	23950,13	11914,80	51608,33	10927,32
Test de Levene (p-valor)		> 0,05		> 0,05		> 0,05	
T de Student (p-valor)		0,6083		0,1631		0,7455	

Tabla 14: resumen de los datos estadísticos obtenidos para la categoría de calidad baja para el tiempo de posedición

En la figura 27 aparecen los resultados del tiempo de posedición representados con los segmentos de calidad baja de menos de cinco palabras. La media de tiempo de posedición en este caso para el motor de TAN es de 15,78 segundos y de 14,23 segundos para el motor de TAE. El test de Levene da un resultado superior al 0,05, con lo cual, la distribución de los datos es probabilística. Tras este, se efectuó el test de Student, que arrojó un resultado de 0,6083. Así, el esfuerzo de posedición es mayor para el motor de TAN que para el de TAE, pero los resultados no son estadísticamente significativos.

Sucede al contrario con los resultados extraídos a partir de los segmentos de seis a diecinueve palabras. La gráfica de la figura 28 vuelve a mostrar valores atípicos, sobre todo en el caso de la muestra de TAE. La media de TAN es de 20,52 segundos, mientras que es de 23,95 para TAE. El test de Levene arrojó un valor superior a 0,05. Por su parte, la prueba de Student efectuada resultó en un valor de 0,1631. Esto muestra que, aunque el tiempo de posedición de la TAN sea inferior, los valores no son estadísticamente significativos.

En el caso de la figura 29, en la gráfica con los resultados de los segmentos de más de veinte palabras, no se muestran variables atípicas. Al contrario que en el caso de los segmentos de seis a diecinueve palabras, aquí el tiempo de posedición de media es mayor para el motor de TAN (54,92) que para el de TAE (51,60). El test de Levene arrojó un valor superior a 0,05, que determina que la distribución de los datos es probabilística. En cuanto al test de Student, presentó un valor de 0,7455. Con lo cual, aunque el tiempo de posedición sea mayor con el motor de TAN, estos valores no son significativos estadísticamente.

De nuevo, para facilitar la comprensión de los resultados, se presentan en forma de tabla a continuación:

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

Tabla 15 resumen de los datos recogidos de la categoría de calidad baja

Al igual que sucedía en la segunda categoría, de calidad intermedia, los tiempos de posedición de la categoría de baja calidad son una muestra demasiado heterogénea. Por ello, los p-valor obtenidos con las distintas pruebas no son significativos. Esto no sucede en los casos de distancia de edición, ya que al tener muestras más homogéneas los resultados sí son estadísticamente significativos.

En ocho de los nueve casos posibles, la distancia de edición es inferior al usar el motor de TAN que en TAE. Solamente en dos ocasiones estos datos no son estadísticamente significativos, a saber: cuando la calidad es alta y los segmentos tienen más de veinte palabras (figura 11) y cuando la calidad es media y los segmentos tienen de seis a diecinueve palabras (figura 13). En lo que respecta al tiempo de posedición, en todos los casos los participantes tardaron más al poseditar los segmentos que provenían de la TAN que de la TAE. No obstante, en estos casos las muestras son demasiado heterogéneas.

5 Conclusiones

Tras haber concluido el apartado anterior de resultados e interpretación de estos, en este nos centraremos en las conclusiones obtenidas. Asimismo, se tratará de justificar los resultados recogidos. Por último, se mencionarán las posibles líneas de investigación a partir de este trabajo y sus resultados.

5.1 MT Ranking

En primer lugar, tal y como se comentaba en el apartado de Objetivos presentado en la introducción, este trabajo tenía por objetivo responder a una serie de cuestiones. Entre ellas, se quería conocer la percepción real de los traductores al trabajar con un motor de traducción automática neuronal y uno de estadística. En concreto, se buscaba saber si los traductores percibían la traducción automática neuronal más productiva, entendiendo la productividad como el menor número de ediciones en el menor tiempo posible. Así, se llevó a cabo una prueba de MT Ranking en DQF con nueve participantes. Los resultados muestran que en un 70 % de las ocasiones los traductores consideran que la traducción automática neuronal es más productiva que la estadística. Estos resultados van de la mano con los obtenidos en las distintas pruebas de evaluación, en las que se evaluaron los segmentos de traducción automática en bruto en términos de fluidez y precisión. En esta prueba se confirma que la traducción automática neuronal recoge una mayor fluidez y precisión, tal y como se muestra en las gráficas presentadas en el apartado de análisis de resultados, en las que la TAN concentra porcentajes más altos en las categorías que corresponden a la mayor calidad.

De esta forma, en ambos casos la traducción automática neuronal se percibe, en términos de productividad (en tiempo y número de ediciones) y calidad, mejor que la estadística. Probablemente esto sea debido al tipo de errores que se dan en los dos motores, dado que, tal y como se ve en las pruebas de calidad, la neuronal se percibe más fluida y más precisa. Sería necesario continuar investigando para comprobar qué sucede realmente. Para ello, sería posible repetir las pruebas de evaluación de calidad con un número mayor de evaluadores, para cotejar los resultados.

En lo que se refiere a las posibles líneas de investigación, en el caso de la prueba de MT Ranking, sería interesante continuar trabajando con los datos recabados en ella. Mediante ellos, se podría confrontar qué segmentos son coincidentes entre todos los participantes y en cuáles algunos de ellos han optado por el motor de traducción automática estadística, en lugar de escoger el neuronal como en el resto de las ocasiones.

5.2 Pruebas de productividad

En segundo lugar, se buscaba observar la productividad de estos dos tipos de motores. Para conseguirlo, se llevaron a cabo distintas pruebas de productividad en DQF. Mediante estas, cabe destacar las siguientes conclusiones:

- La distancia de edición es inferior en los segmentos poseditados con el motor de traducción automática neuronal que en los segmentos poseditados con el motor de traducción automática estadística.
- El esfuerzo de posedición, en tiempo, es mayor en los segmentos poseditados con el motor de traducción automática neuronal que en los segmentos poseditados con el motor de traducción automática estadística.

Por lo tanto, si bien es cierto que la traducción automática neuronal requiere, por norma general, un número menor de ediciones que al usar un motor estadístico, al usar el motor de traducción automática neuronal el participante, de nuevo por norma general, emplea un mayor tiempo de posedición. Una explicación que justifique este fenómeno podría ser que, puesto que el traductor percibe la traducción automática neuronal como más productiva, al no detectar los errores con facilidad, a simple vista, tarda más en encontrarlos, lo que le lleva más tiempo para poseditar los segmentos.

Para continuar con esta línea de investigación, se podrían realizar de nuevo las pruebas, registrándolas mediante un *eye-tracker*. De esta forma se vería exactamente dónde focaliza el participante su atención. No obstante, emplear un sistema de seguimiento ocular hace que el experimento se deba realizar con un número menor de sujetos y, de esa forma, se perdería información relevante. Asimismo, en estos casos se

vio que la muestra de datos era demasiado heterogénea. Por lo tanto, sería necesario repetir las pruebas con otro grupo de participantes en un entorno controlado.

Asimismo, otra línea de investigación relevante podría ser trabajar en determinar qué errores se han producido en cada tipo de motor, para mejorar la enseñanza de la posedición en el ámbito académico. Para ello, nos valdríamos de los segmentos poseditados y se establecerían los errores en las distintas categorías de MQM.

Es necesario realizar un aparte para las pruebas de calidad. Si bien estas pruebas eran un instrumento para abarcar un mayor número de resultados, de estas también se pueden extraer conclusiones. Tal y como se mencionaba en el apartado de Objetivos, se buscaba en este trabajo resolver la hipótesis de si el motor de traducción automática neuronal era más fluido, pero menos preciso que el motor de traducción automática estadística. Como se ha podido ver, los resultados obtenidos en las pruebas indican que el motor de traducción automática neuronal es más fluido y más preciso que el de estadística. Esto puede deberse a que, tal y como sucede con los resultados de la prueba de MT Ranking, el traductor/poseditor percibe los resultados del motor de traducción automática neuronal como más adecuados, en términos de fluidez y precisión, aunque le lleve más tiempo poseditarlos. Se podría continuar investigando esta línea mediante la repetición de las pruebas con una tipología textual distinta.

6 Bibliografía

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate, 1–15. Recuperado de http://doi.org/10.1146/annurev.neuro.26.041002.131047.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study, 2. Recuperado de http://doi.org/10.18653/v1/D16-1025.

Casacuberta Nolla, F., & Peris Abril, Á. (2017). Traducció automàtica neuronal. *Tradumàtica: tecnologies de la traducció*, *15*, 66-74. Recuperado de https://doi.org/10.5565/rev/tradumatica.203.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1). Recuperado de https://doi.org/10.1515/pralin-2017-0013.

Chan, S.-W. (Ed.). (2014). *Routledge Encyclopedia of Translation Technology*. Londres: Routledge.

Chéragui, M. A. (2012). Theoretical Overview of Machine Translation. *Proceedings ICWIT*, 160-168. Recuperado de citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.1463&rep=rep1&type=pdf#page=176.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. Recuperado de http://doi.org/10.3115/v1/W14-4012.

Chunyu, K., & Tak-ming, B. W. (2015). Evaluation in machine translation and computer-aided translation. *The Routledge encyclopedia of translation technology*. Londres: Routledge.

Dale, R. (Ed.), Moisl, H. (Ed.), Somers, H. (Ed.). (2000). *Handbook of Natural Language Processing*. Boca Ratón: CRC Press.

Densmer, L. (2014). Light and Full MT Post-Editing Explained. Moravia.com. Recuperado de https://info.moravia.com/blog/bid/353532/Light-and-Full-MT-Post-Editing-Explained.

Doherty, S., & Gaspari, F. (2013). Understanding and Implementing Effective Translation Quality Evaluation Techniques. *Centre for Next Generation Localisation*. Dublín: Dublin City University. Recuperado de http://www.qt21.eu/launchpad/sites/default/files/QTLP%20GALA%20Webinar%203.pdf.

Dorr, B., Olive, J., McCary, J., & Christianson, C. (2011). Machine Translation Evaluation and Optimization. *Handbook of Natural Language Processing and Machine Translation*. Recuperado de https://doi.org/10.1007/978-1-4419-7713-7_5.

Fernández-Rodríguez, M. (2010). Evolución de la traducción asistida por ordenador: de las herramientas de apoyo a las memorias de traducción. Sendebar, 21, 201–230. Recuperado de http://revistaseug.ugr.es/index.php/sendebar/article/view/374.

Fields, P., Hague, D. R., Koby, G. S., Lommel, A., & Melby, A. (2014). What Is Quality? A Management Discipline and the Translation Industry Get Acquainted. *Tradumàtica*, *12*, 404-412. Recuperado de https://revistes.uab.cat/tradumatica/article/download/75/pdf.

Ginestí, M., & Forcada, M. L. (2009). La traducció automàtica en la pràctica: aplicacions, dificultats i estratègies de desenvolupament. *Caplletra: Revista Internacional de Filologia*, *46*, 43-60. Recuperado de http://roderic.uv.es/handle/10550/48542.

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

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

Görög, A. (2014). Quality evaluation today: the Dynamic Quality Framework. *Proceedings of Translating and the computer*, *36*, 155-164. The International Association for Advancement in Language Technology. Recuperado de http://www.tradulex.com/varia/TC36-london2014.pdf.

Hutchins, W. J. (1986). *Machine Translation: Past, Present and Future*. Chichester: Ellis Horwood.

Hutchins, W. J. & Somers, H. L. (1992). *An Introduction to Machine Translation*, Londres: Academic Press.

Hutchins, W. J. (1995). Machine Translation: A Brief History. *Concise history of the language sciences: from the Sumerians to the cognitivists*, 431-445 Recuperado de https://pdfs.semanticscholar.org/330f/7f1357b640671c4c88c05f9c90c6ccf0de46.pdf.

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2014). On Using Very Large Target Vocabulary for Neural Machine Translation. Recuperado de http://doi.org/10.3115/v1/P15-1001.

Koby, G. S., Fields, P., Hague, D. R., Lommel, A., & Melby, A. (2014). Defining Translation Quality. *Tradumàtica*, *12*, 413-420. Recuperado de http://revistes.uab.cat/ojs-tradumatica/tradumatica/issue/view/5.

Kuang, S., & Xiong, D. (2016). Automatic Long Sentence Segmentation for Neural Machine Translation. En C.-Y. Lin, N. Xue, D. Zhao, X. Huang, & Y. Feng (Eds.), *Natural Language Understanding and Intelligent Applications* (pp. 162–174). Cham: Springer International Publishing.

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M. & Volk, M. (2013). Assessing post-editing efficiency in a realistic translation environment. *Proceedings of Workshop on Post-editing Technology and Practice*, 83–91. Recuperado de http://www.mt-archive.info/10/MTS-2013-W4-Laubli.pdf.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica*, 12, 455-463. Recuperado de https://ddd.uab.cat/record/130144/.

Luong, M., Sutskever, I., Le, Q. V, Vinyals, O., & Zaremba, W. (2014). Addressing the Rare Word Problem in Neural Machine Translation. Recuperado de http://doi.org/10.3115/v1/P15-1002.

Luong, M.-T., & Manning, C. D. (2015). Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. Recuperado de http://arxiv.org/abs/1604.00788.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *EMNLP*, *11*, 1412-1421. Recuperado de http://doi.org/10.18653/v1/D15-1166.

Marie, B., & Chesnay, L. (2015). Touch-Based Pre-Post-Editing of Machine Translation Output. *Emnlp*, 1040–1045. Recuperado de http://aclweb.org/anthology/D/D15/D15-1120.pdf.

Martín-Mor, A., Piqué, R., & Sánchez-Gijón, P. (2016). *Tradumàtica: tecnologies de la traducció*. Vic: Eumo.

Martín-Mor, A., Doȋru, G., & Ortiz, S. (2017). MTradum¨tica: Free Statistical Machine Translation Customisation for Translators. *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, 65-67. Recuperado de https://ddd.uab.cat/record/174910.

Martínez-Mateo, R. (2014). A deeper look into metrics for Translation Quality Assessment (TQA): A case study. *Miscelanea: A Journal of English and American Studies*, 49, 73-93. Recuperado de http://www.aclweb.org/anthology/W15-4910.

Moorkens, J., & O'Brien, S. (2015). Post-Editing Evaluations: Trade-offs between Novice and Profession al Participants. *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, 75-81. Recuperado de http://www.aclweb.org/anthology/W15-4910.

Moorkens, J., O'Brien, S., Silva, I. A. L., Fonseca, N., & Alves, F. (2015). Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29 (3-4), 267-284.

Moorkens, J., & O'Brien, S. (2016). Assessing User Interface Needs of Post-Editors of Machine Translation. *Human Issues in Translation Technology: The IATIS Yearbook*. Abingdon: Routledge. Recuperado de https://www.researchgate.net/publication/312121765/download.

Moorkens, J. (2017). Under pressure: translation in times of austerity. *Perspectives*, 464-477. Recuperado de https://www.tandfonline.com/doi/abs/10.1080/0907676X.2017.1285331.

Multidimensional Quality Metrics Definition. (s. f.). Recuperado de http://www.qt21.eu/mqm-definition/definition-2015-12-30.html.

Nirenburg, S., Somers, H. L., & Wilks, Y. (2003). *Readings in Machine Translation*. Cambridge, Mass.: MIT Press.

Norma Internacional ISO/DIS 18587.2, Translation services — Post-editing of machine translation output — Requirements. Recuperado de https://www.iso.org/obp/ui/#iso:std:iso:18587:dis:ed-1:v2:en:term:2.4.5.

Norma Internacional ISO/TS 11669:2012 — Translation projects — General guidance. Recuperado de https://www.iso.org/standard/50687.html.

O'Brien, S. (2006). Machine-Translatability and Post-Editing Effort: An Empirical Study Using Translog and Choice Network Analysis. Dublín: ADAPT/Dublin City University. Recuperado de http://doras.dcu.ie/18118/1/Sharon_O%27BrienV1.pdf.

O'Brien, S., Choudhury, R., Van der Meer, J., Aranberri Monasterio, N. (2011) TAUS Dynamic Quality Evaluation Framework: TAUS Labs report, TAUS.net. https://www.taus.net/file-downloads/download?path=Reports%252FFree%2BReports%252Ftausdynamicquality.pdf.

O'Brien, S. (2012). Towards a Dynamic Quality Evaluation Model for Translation. *Journal of Specialised Translation*, *17*, 55-77. Recuperado de http://www.jostrans.org/issue17/art_obrien.php.

O'Brien, S. (2014). Translation Quality - It's time that we agree. TAUS.com Recuperado de https://www.taus.net/think-tank/articles/event-articles/translation-quality-it-s-time-that-we-agree.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia: Association for Computational Linguistics. Recuperado de https://doi.org/10.3115/1073083.1073135.

Presas, M., Cid-Leal, P., & Torres-Hostench, O. (2016). Machine translation implementation among language service providers in Spain: A mixed methods study. *Journal of research design and statistics in linguistics and communication science*, *3*(1), 126-144.

Pujadas, E. C. (2015). Automatic Machine Translation Evaluation: A Qualitative Approach. Barcelona: Universitat de Barcelona. Recuperado de http://diposit.ub.edu/dspace/bitstream/2445/65906/1/ECP_PhD_THESIS.pdf.

Rico Pérez, C. (2017). Training translators in Machine Translation. *Tradumàtica: tecnologies de la traducció*, *15*, 75-96. https://doi.org/10.5565/rev/tradumatica.200.

Torrejón, E., & Rico, C. (2012). Skills and Profile of the New Role of the Translator as MT Post-editor. *Tradumàtica: tecnologies de la traducció*, *10*, 166–178. Recuperado de http://revistes.uab.cat/tradumatica/article/view/18.

Sánchez-Gijón, P. (2016). La posedición: hacia una definición competencial del perfil y una descripción multidimensional del fenómeno. *Sendebar: Revista de la Facultad de Traducción e Interpretación*, *27*, 151-162. Recuperado de http://revistaseug.ugr.es/index.php/sendebar/article/view/4016.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural Machine Translation of Rare Words with Subword Units. Recuperado de http://doi.org/10.18653/v1/P16-1162.

Systran. (2016). Systran White Paper. Systransoft.com. Recuperado de http://www.systransoft.com/download/white-papers/systran-white-paper-PNMT-12-2016_2.pdf.

TAUS. (2010). MT Post-editing Guidelines. TAUS.com. Recuperado de https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines.

Torres Hostench, O., Cid-Leal, P., Presas, M., Piqué Huerta, R., Sánchez-Gijón, P., Aguilar Amat, A., … Ángel, M. (2016). El uso de traducción automática y posedición en las empresas de servicios lingüísticos españolas: informe de investigación ProjecTA 2015. Recuperado de https://ddd.uab.cat/record/148361.

Tripathi, S., & Sarkhel, J. K. (2010). Approaches to machine translation. *International journal of Annals of Library and Information Studies*, *57*, pp. 388-393. Recuperado de https://www.researchgate.net/publication/228574546_Approaches_to_machine_translation.

Turian, J. P., Shea, L., & Melamed, I. D. (2006). *Evaluation of Machine Translation and its Evaluation:* Fort Belvoir: Defense Technical Information Center. https://doi.org/10.21236/ADA453509.

Vasconcellos, M. (1988). *Technology as Translation Strategy*. Philadelphia: John Benjamins Publishing.

Vasconcellos, M. (1991). Machine translation and the language barrier. Recuperado de http://mt-archive.info/Vasconcellos-1991.pdf.

Vilar, D., Xu, J., D'Haro, L. F., & Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. En *Proceedings of LREC*, 697–702. Recuperado de http://hnk.ffzg.hr/bibl/lrec2006/pdf/413_pdf.pdf.

Woǧk, K., & Korǵinek, D. (2017). Comparison and Adaptation of Automatic Evaluation Metrics for Quality Assessment of Re-Speaking. *Computer Science*, *18*(2), 129. Recuperado de https://doi.org/10.7494/csci.2017.18.2.129.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V, Norouzi, M., Macherey, W., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv E-Prints, 1–23. Recuperado de http://doi.org/10.1038/nrn2258.

Yamagata Europe. (2018) Is neural machine translation always the best option? - Yamagata Europe. Recuperado de https://www.yamagata-europe.com/en-gb/blog/is-neural-machine-translation-always-the-best-option.

Zhechev, V. (2012). Machine Translation Infrastructure and Post-editing Performance at Autodesk. *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, 87–96. Recuperado de https://pdfs.semanticscholar.org/6516/04d416f85963082b02c5af2fe1c1ccdbf215.pdf.

7 Anexos

Enlace disponible de consulta a la hoja de datos preparada a raíz de las evaluaciones de calidad (1): http://bit.ly/datos_fluidez_precision_TFM.

Enlace disponible de consulta a la hoja de datos preparada a raíz de las pruebas de productividad (2): http://bit.ly/datos_productividad_participantes_TFM.

Enlace disponible de consulta a la hoja de datos preparada a raíz de las pruebas de productividad: http://bit.ly/datos_productividad_generales_TFM.