Máster en Tradumática: Tecnologías de la Traducción

EVALUACIÓN COMPARATIVA DE CORPUS USADOS EN EL ENTRENAMIENTO DE UN MOTOR DE TAE PARA LA TRADUCCIÓN DE GUÍAS DOCENTES DE LA UNIVERSIDAD AUTÓNOMA DE BARCELONA

TRABAJO DE FIN DE MÁSTER

Curso 2019-2020

Lucía Carolina Pacheco Ortiz

Tutor: Ramón Piqué Huerta

Facultad de Traducción e Interpretación, UAB

Universitat Autònoma de Barcelona

RESUMEN

En este trabajo de fin de máster se lleva a cabo la creación de seis motores de traducción automática estadística (EN>ES) mediante la plataforma de MTradumática, entrenados con corpus de entrenamiento y optimización reunidos a partir de guías docentes de programas del área de Humanidades de la Universidad Autónoma de Barcelona. Se realiza una comparación entre las traducciones realizadas con dichos motores, así como entre ellas y las traducciones creadas con el motor de traducción automática Google Translate, para lo cual se usa el sistema de evaluación BLEU. Así se pretende establecer si un motor de traducción automática estadística personalizado ofrece mejores resultados a la hora de traducir las guías docentes de la UAB que un motor de traducción automática neuronal de carácter general tal como Google Translate, e igualmente se pretende discernir cuál es el mejor de los motores especializados y qué características deben cumplir un corpus de entrenamiento y un corpus de optimización.

Palabras clave: *traducción automática, BLEU, motor de traducción automática, MTradumática, corpus de entrenamiento, corpus de optimización*

ABSTRACT

For this Master's Degree Dissertation, six statistical machine translation engines (EN>ES) were created using the MTradumática platform, trained with training and optimization corpus gathered from syllabi for programs in the Humanities area of the Universidad Autónoma de Barcelona. A comparison is made between the translations created with those engines, as well as between them and the translations created with the Google Translate machine translation engine; using the BLEU evaluation system. The aim is to establish whether a customized statistical machine translation engine offers better results when translating the UAB syllabi than a general neural machine translation engine such as Google Translate, and also to discern which is the best of the specialized engines, and what features should a training corpus and an optimization corpus have.

Keywords: *machine translation, BLEU, machine translation engine, MTradumática, training corpus, optimization corpus*

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

La traducción automática es cada vez una parte más integral de la profesión del traductor, así como de la vida diaria de las personas, empresas y organizaciones. Actualmente es posible, no sólo utilizar herramientas de traducción automática sofisticadas y disponibles para el público en general, sino crear nuestros propios motores especializados de acuerdo con la necesidad específica que precisamos solventar. La Universidad Autónoma de Barcelona ofrece una plataforma de código abierto para este fin, llamada MTradumática.

En este trabajo se pretende realizar una comparación entre seis motores de traducción automática personalizados creados con la plataforma de MTradumática, así como entre ellos y el motor de traducción automática Google Translate. De esta manera se podrá establecer no sólo la diferencia de calidad entre los distintos motores, sino discernir las características ideales de un corpus de entrenamiento y un corpus de optimización a la hora de crear un motor de traducción automática especializado.

2. OBJETIVOS

2.1. Objetivo General

Evaluar la calidad de diferentes motores de traducción automática entrenados con el programa MTradumática para la traducción de guías docentes de español a inglés.

2.2. Objetivos Específicos:

2.2.1. Crear seis motores de traducción automática a partir de corpus basados en guías docentes de la facultad de Humanidades de la Universidad Autónoma de Barcelona, utilizando el programa de entrenamiento MTradumática

2.2.2. Traducir cinco guías docentes utilizando los motores de traducción entrenados con MTradumática

2.2.3. Comparar las traducciones de las cinco guías docentes utilizando el sistema de evaluación BLEU.

3. PREGUNTA E HIPÓTESIS

3.1. **Pregunta**:

3.1.1. ¿Existe una correlación entre la limpieza de los corpus de entrenamiento y optimización, y la calidad de las traducciones realizadas con los motores entrenados con el programa MTradumática, según la puntuación BLEU?

3.1.2. ¿Alguno de los motores alcanza un nivel de calidad adecuado para su uso de acuerdo con el método de calidad BLEU?

3.1.3. ¿Hay factores adicionales que afecten la calidad de la traducción hecha por los motores de traducción entrenados con el programa MTradumática?

3.2. Hipótesis:

La limpieza de los corpus de entrenamiento y optimización tiene una correlación positiva directa con la calidad de traducción de los motores de traducción automática (TAE) entrenados con el programa MTradumática, y los motores entrenados con los corpus limpios alcanzan un nivel de calidad fluida que permite su uso para la traducción automática de guías docentes.

4. MARCO TEÓRICO

El marco teórico de este trabajo comienza con una introducción a los conceptos básicos de la traducción automática, y posteriormente, explorará los programas de entrenamiento de motores de traducción automática, específicamente el programa MTradumática. Después tratará el concepto de Posedición, y luego se centrará en la evaluación de calidad, y específicamente en el sistema de evaluación BLEU.

4.1. Conceptos básicos de la traducción automática

La *European Association for Machine Translation* (EAMT) define la traducción automática como "la aplicación de los computadores a la tarea de traducir textos de un lenguaje natural a otro"(*EAMT | European Association for Machine Translation*, s. f.)

Existen dos tipos de motores de traducción automática según la arquitectura de su sistema: los motores de traducción basados en reglas, y los motores de traducción basados en corpus (Ping, 2009). Estos últimos se dividen en tres enfoques diferentes: motores de traducción basados en ejemplos, motores de traducción estadística y motores de traducción neuronal (Martín-Mor, 2017). De igual manera, existen motores de traducción automática que aplican un enfoque híbrido.

4.1.1. Motores de traducción basados en reglas (TABR)

Son los primeros motores de traducción automática que aparecieron en el panorama; y dependen de datos lingüísticos como diccionarios, morfológicos y bilingües, gramáticas, y archivos de reglas de transferencia estructural. escritas específicamente para cada combinación de idiomas. El sistema de TABR genera las oraciones después de pasar a través de procesos de comprobación léxica, morfológica y sintáctica, creados por un humano. Por esta razón requiere un gran esfuerzo de desarrollo, pero funciona bien entre lenguas cercanas y con pocos recursos. (Forcada, 2009).

4.1.2. Motores de traducción basados en ejemplos

El motor de traducción basado en ejemplos, también llamado motor de traducción por analogía, fue propuesto por Nagao en el artículo *A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle* (1984).

> Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases (very often, into case frame units), then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference (Nagao, 1984)

Según Nagao, la traducción basada en ejemplos tiene tres pasos: el emparejamiento de fragmentos comparándolos con una gran base de datos de ejemplos de la vida real, la identificación de los fragmentos de traducción correspondientes, y la combinación de dichos fragmentos para crear un texto

de llegada. Es decir, dados uno o más textos paralelos, el sistema analiza la frase a traducir, la divide en segmentos más pequeños cuyas traducciones son recuperadas de los textos paralelos y las combina para producir una nueva traducción en lengua meta (Sánchez-Martínez, 2012).

4.1.3. Motores de traducción estadística (TAE)

A partir de los años noventa, surgieron programas que "aprenden a traducir" a partir de corpus de miles frases (Forcada, 2009). Al usar la traducción automática estadística (TAE), las palabras y las frases (secuencias de palabras) alineadas son la base de un modelo de traducción que usa las frecuencias palabra-palabra y frase-frase. Ya que el motor de traducción se basa en corpus, crear un corpus de texto bilingüe de alta calidad es esencial para el éxito del motor TAE. Si este es el caso, se pueden lograr resultados impresionantes al traducir textos similares a aquellos que se encuentran en el corpus de entrenamiento (Ping, 2009).

A diferencia de los motores de TABR, un motor de TAE se puede crear en un espacio de tiempo relativamente corto. La mayoría del tiempo de trabajo se emplea en la limpieza del corpus pero, una vez está listo, el proceso de entrenamiento se puede llevar a cabo en unos días. Varios motores de TAE usan lenguas puente para aumentar el número de combinaciones de idiomas disponible (Martín-Mor, 2017).

En el artículo *Statistical Machine Translation: A Guide for Linguists and Translators* (2011), Mary Hearne y Andy Way señalan que la TAE emplea dos procesos distintos: *entrenamiento* y *decodificación*, y los resumen de la siguiente manera:

> The translation model effectively comprises a bilingual dictionary where each possible translation for a given source word or phrase

has a probability associated with it. However, the model does not resemble a conventional dictionary where plausible entries only are permitted; many of the entries represent translations that are unlikely but not impossible, and the associated probabilities reflect this. The language model comprises a database of target-language word sequences (usually ranging between 1 and 7 words in length), each of which is also associated with a probability (…) These induced models are then used during decoding, the process which actually yields a translation. The decoding process essentially treats translation as a search problem: given the sentence to be translated, search over all possible translations permitted by the translation model, and all possible reorderings thereof, for the one which is assigned the highest overall probability according to the translation and language models (Hearne & Way, 2011).

4.1.4. Motores de traducción híbridos

Los motores de traducción híbridos son sistemas que combinan los métodos estadísticos de la TAE o de la traducción basada en ejemplos con algunos métodos de la traducción basada en reglas (Hutchins, s. f.).

4.1.5. Motores de traducción neuronal

Forcada (2017) define la traducción automática neuronal (TAN) como una nueva generación de traducción basada en corpus, similar a la tecnología de TAE en el sentido de ser entrenada usando un gran corpus que consiste en inmensas memorias de traducción que contienen cientos de miles, o incluso millones, de unidades de traducción; pero usando un enfoque informático completamente diferente llamado redes neuronales.

Parra (2018) describe estas redes de la siguiente manera:

> estas redes pretenden emular la manera en la que funcionan las neuronas en nuestro cerebro (…) Así, utilizando técnicas de aprendizaje automático, el ordenador aprende a traducir a partir de grandes cantidades de textos paralelos que además incluyen todo tipo de información lingüística y no lingüística (…) Gracias a la manera de relacionar la información asociada a cada palabra y a la de las palabras de una frase, el ordenador es capaz de aprender a traducir de una manera más eficiente.

Forcada menciona que los sistemas TAN requieren usualmente corpus muy grandes, y que su entrenamiento es informáticamente muy demandante, de tal manera que el proceso de entrenamiento típicamente puede durar entre días y meses lo que los hace más difíciles de entrenar que los sistemas TAE (2017). Cabe también señalar que es complejo poseditar traducciones producidas por sistemas TAN, ya que estos sistemas suelen generar traducciones que en un primer momento parecen más correctas pero que en realidad no siempre lo son. Esto es porque generalmente las frases son gramaticales y tienen sentido, por lo que un cambio de significado puede pasar desapercibido si la frase en sí es posible dentro del contexto en el que se encuentra. También se ha demostrado que en el caso de frases muy largas (de más de 25-30 palabras) los sistemas basados en redes neuronales suelen obtener peores resultados que los sistemas estadísticos (Parra, 2018).

A otro nivel, los sistemas de TA se pueden clasificar en genéricos o específicos (Martín Mor, 2017).

4.1.6. Motores de TA genéricos

Los sistemas genéricos de TA son sistemas de uso general que traducen textos en cualquier área o campo del conocimiento, y pueden usarse, por

ejemplo, para entender la esencia de la información contenida en una página web en un idioma extranjero (Ping, 2009) Los motores de traducción automática más conocidos como Google Translate, Microsoft Translate y Deepl, entre otros, son ejemplos de motores genéricos.

4.1.7. Motores de TA personalizados o específicos

Los sistemas de TA personalizados o específicos se dirigen a grupos de usuarios que trabajan en áreas o campos específicos, y son mucho más efectivos que los sistemas de TA genéricos (Ping, 2009). Entre más cerca se encuentren los datos que se usan (para entrenar un motor de traducción automática) al tipo de datos que se quiera traducir, mejor serán los resultados (Koehn, 2020).

4.2. Entrenamiento de motores TA

Dogru et al. Afirman que está ampliamente aceptado que la TA funciona mejor con corpus paralelos de un campo específico, y que en los enfoques de TA basados en corpus, entre más específico el campo de entrenamiento, mejor será el resultado de traducción (2018). De ahí la importancia de la existencia de herramientas que nos permitan entrenar motores de TA según nuestras propias necesidades y el tipo de textos que vayamos a traducir (Fernández Ruiz & Sánchez-Gijón, 2019).

Existen herramientas de entrenamiento de TA (específicamente TAE) de sistema propietario como Microsoft Translator Hub y LetsMT, y de software libre como Moses, lanzado bajo una Licencia Pública General Reducida de GNU (LGPL por sus siglas en inglés) (Martín-Mor, 2017).

La existencia de programas de TA libres/de código fuente abierto (como Moses) ha favorecido la adopción de sistemas de TA sin necesidad de invertir grandes cantidades de dinero en implementar técnicas y métodos de TA que han demostrado su utilidad en el ámbito experimental. La gran mayoría de estos programas han sido

desarrollados por la comunidad científica y se encuentran en constante desarrollo, incorporando los últimos avances científicos (Sánchez-Martínez, 2012). Algunas plataformas derivadas de este código abierto son la plataforma comercial KantanMT (KantanMT, 2015), y MTradumática (Martin-Mor & Piqué i Huerta, 2017), en la cual se ahondará a continuación.

4.2.1. MTradumática

MTradumática es el entrenador de motores de traducción automática estadística creado por el grupo de investigación Tradumática de la Universidad Autónoma de Barcelona (Martin-Mor & Piqué i Huerta, 2017) . MTradumática es una plataforma gratuita basada en Moses para entrenar y usar motores de traducción automática estadística con una interfaz gráfica fácil de usar, cuyo objetivo es ofrecer a los traductores una herramienta gratuita para personalizar sus propios motores de traducción automática estadística y mejorar su productividad (*What Is MTradumàtica?*, s. f.).

El proceso de entrenamiento sigue las siguientes etapas: preparación de la documentación, la preparación de recursos para el sistema de traducción automática, el entrenamiento del sistema, la preedición de textos, la traducción automática, la posedición y la retroalimentación del sistema (Martin-Mor & Piqué i Huerta, 2017).

Para entrenar un motor de traducción automática en MTradumática, se deben seguir los siguientes pasos (*What Is MTradumàtica?*, s. f.):

1) Subir a la plataforma archivos en formatos de textos paralelos frase por frase tales como Moses o TMX
2) Crear y gestionar monotextos a partir de dichos archivos
3) Crear los modelos de lengua a partir de los monotextos

4) Construir y gestionar bitextos
5) Entrenar modelos de traducción automática estadística usando los bitextos y los modelos de lengua

A partir de ahí, el traductor podrá traducir textos cortos y documentos usando estos motores, así como inspeccionarlos y evaluarlos.

4.3. Edición

En la entrevista *Ray Kurzweil on Translation Technology* (Kelly, 2011), Kurzweil describió la traducción como "el tipo de trabajo de más alto nivel se uno se pueda imaginar", y agregó, "el epítome de la inteligencia humana es nuestra habilidad para dominar el lenguaje. Por eso Alan Turing basó el test de Turing, que es una prueba de si un computador opera al nivel de un ser humano o no, en el dominio del lenguaje".

La TA completamente automatizada es casi gratis, pero el resultado, aunque sirve para hacerse una idea de la traducción, es muy inferior a la calidad de los traductores humanos expertos. Sin embargo, los beneficios ostensibles en tiempo y costo de la TA son demasiado atractivos para resistir (Green et al., 2013). Por esto, surge la necesidad de contar con editores, a quienes Bar-Hillel (1951) se refiere como los "socios humanos" de la máquina, es decir, del programa de traducción automática. Dichos socios pueden desarrollar tareas de preedición, posedición, o ambas.

4.3.1. Preedición

La preedición es la preparación de un texto para su traducción por un sistema de traducción automática o semi-automática (motor de traducción automática, memoria de traducción) (*Translation services - Service requirements*, s. f.). Según Bar-Hillel (1951), la preedición consiste en eliminar

las ambigüedades morfológicas y sintácticas y reorganizar el texto de partida de acuerdo con el estándar de la lengua de llegada, siguiendo un conjunto de instrucciones disponible en su propio idioma.

Según Pym, la preedición se basa en la revisión del texto original para remover “indicadores de negativos de traducibilidad” o elementos que probablemente sean problemáticos para la traducción automática. La relación costo-beneficio de la preedición con respecto a la Posedición sólo se evidencia cuando un texto dado se va a traducir a más de un idioma de llegada, y aumenta aritméticamente con cada lengua de llegada adicional, dependiendo de qué tan específicos sean los indicadores de traducibilidad para cada combinación de idiomas (2019).

4.3.2. Posedición

El principal rol del poseditor consiste en eliminar ambigüedades semánticas, en adición al perfeccionamiento estilístico (Bar-Hillel, 1951). En la mayoría de contextos profesionales, el resultado de los sistemas de traducción automática necesita ser poseditado para alcanzar los estándares de calidad deseados (Nunes Vieira, 2019).

Sánchez-Gijón (2016) señala que la posedición (PE), entendida como la edición de segmentos obtenidos mediante traducción automática (TA), irrumpió con fuerza en la industria de la traducción hace algunos años y que desde entonces, el uso de la traducción automática como recurso para traducir ha crecido, demostrando así que la fórmula de TA + PE ha llegado para quedarse. Aranberri afirma que esto no sólo se debe a que la calidad de los sistemas punteros haya mejorado considerablemente, sino también a que las empresas del sector han reconocido que incluso una TA imperfecta

puede ser útil para satisfacer las demandas actuales del mercado de la traducción (Aranberri, 2014).

La posedición ha crecido al punto de que en el año 2017, la Organización Internacional de Normalización (ISO) publicó un estándar para ella, distinto del estándar para la traducción, a través de la norma ISO:18587 (*ISO 18587:2017(en), Translation services — Post-editing of machine translation output — Requirements*, 2017). En ella se señala que aunque la TA es una solución viable para proyectos de traducción que deben completarse dentro de un plazo muy estrecho o con un presupuesto limitado, no existe un sistema de TA que dé un resultado que se pueda calificar como igual al resultado de una traducción humana y que, por lo tanto, la calidad final de una traducción aún depende de los traductores humanos y su competencia en posedición.

La posedición puede ser de dos tipos: parcial (light post-editing), o compleja (full post-editing). La primera consiste en realizar los cambios necesarios e imprescindibles para que un texto pueda ser comprendido (es decir, el texto final puede contener errores siempre y cuando éstos permitan que el mensaje se transmita satisfactoriamente); mientras que la segunda tiene como objetivo eliminar todo error de la TA y conseguir una traducción de alto nivel, a la par de la traducción tradicional manual, siendo correcto desde una perspectiva gramatical y terminológica, así como estilística, y gozar de una fluidez nativa (Aranberri, 2014).

4.4. TA y Calidad

Melby (Melby, 2012) define qué es una traducción de calidad desde una perspectiva funcional:

> A quality translation demonstrates the levels of accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account both requester and end user needs.

Aunque el área de la evaluación de la TA empezó como un subcampo dentro del desarrollo de la TA, ahora se ha convertido en un campo con un conjunto más amplio de metas e interesados, que deben tomar decisiones sobre si los sistemas de TA pueden suplir sus necesidades (Babych, 2014).

Por otro lado, la *Translation Automation User Society* (TAUS), enfocándose en la calidad de la traducción final antes que en el método de posedición, distingue entre la calidad "suficientemente buena", que define como comprensible y precisa pero no es buena estilísticamente; y la calidad similar o igual a la de una traducción humana, que además de ser comprensible y precisa es estilísticamente correcta, aunque puede no ser tan buena como la alcanzada por un traductor que es hablante nativo (*Pautas para la Posedición de la traducción automática*, 2013).

Actualmente, los métodos y herramientas de evaluación de TA más usados son los de evaluación automatizada tales como BLEU y METEOR. Su principal objetivo es computar puntajes numéricos que caracterizan la "calidad", o el nivel de desempeño, de sistemas específicos de TA. Se espera que estos puntajes se correlacionen con los juicios intuitivos de los humanos sobre ciertos aspectos de la calidad de la traducción, o con ciertas características de uso para los textos traducidos (Babych, 2014).

4.4.1. BLEU

> Papineni, Roukos, Ward y Zhu (2002) describen el sistema de evaluación BLEU como "una evaluación económica, independiente del idioma, y que se

correlaciona altamente con la evaluación humana". Su filosofía es que una traducción automática es mejor entre más cerca se encuentre a una traducción humana profesional. Este sistema requiere 1. Una medida numérica de "similitud de traducciones" y 2. Un corpus de traducciones humanas de referencia de buena calidad.

La métrica de BLEU se encuentra en un rango de 0 a 1 (o de 0 a 100) (Papineni et al., 2002). Una traducción sólo puede obtener un puntaje de 1 (o 100) si es idéntica a una traducción de referencia. Por esta razón, ni siquiera una traducción humana obtendría necesariamente un puntaje de 1.

Según Vashee (*Understanding MT Quality*, s. f.), aunque BLEU como métrica tiene muchas debilidades, es probablemente la métrica de evaluación de calidad más usada en los últimos 15 años, aún durante el apogeo de la traducción automática neuronal. Vashee señala también que para obtener una medición significativa se recomienda usar un corpus de entrenamiento de al menos 1000 frases, ya que un corpus demasiado pequeño puede hacer que dos o tres frases que no encajen bien influencien demasiado el puntaje.

La siguiente es la interpretación de puntuaciones BLEU expresadas como porcentajes (*Evalúa modelos | Documentación de AutoML Translation*, s. f.)

Puntuación BLEU	Interpretación
< 10	Casi inútil
10 - 19	Difícil de captar la esencia
20 - 29	La esencia es clara, pero tiene errores gramaticales significativos

Puntuación BLEU	Interpretación
30 - 40	Comprensible por buenas traducciones
40 - 50	Traducciones de alta calidad
50 - 60	Traducciones de calidad muy alta, adecuadas y fluidas
> 60	Calidad que suele ser mejor que la humana

Tabla 1. Escala de puntajes de la métrica BLEU.

5. METODOLOGÍA

En este capítulo se explicará el proceso de preparación y creación de cuatro motores de traducción automática estadística por medio de la plataforma MTradumática, así como la traducción de cinco guías docentes utilizando y evaluación de estos usando el sistema de evaluación BLEU.

5.1. Preparación del corpus

5.1.1. Descripción del corpus

Este trabajo cuenta con dos corpus: el corpus de entrenamiento y el corpus de optimización. Los segmentos tanto de los corpus como las guías de prueba están en español y tienen sus correspondientes traducciones en inglés.

El corpus de entrenamiento consiste en 97.558 segmentos de diversas longitudes, en español y traducidos al inglés, sacados de guías docentes de materias que forman parte del currículo de varios programas de la facultad de Humanidades de la Universidad Autónoma de Barcelona. Estos segmentos contienen en total 1’306.389 palabras en español y 1’197.157 palabras en

inglés, respectivamente, de acuerdo con el programa Microsoft Word. Este corpus fue sujeto a un proceso de limpieza (véase apartado 5.1.2.) que redujo el número de segmentos a 76.717, y el número de palabras a 1’154.866 en español y 1’023.379 en inglés.

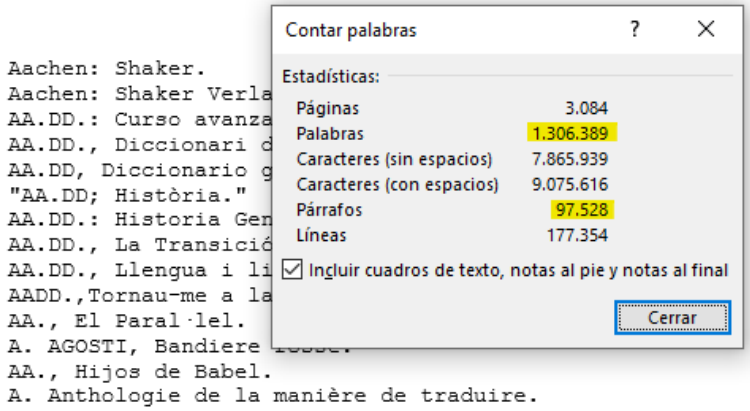


Figura 1. Conteo de palabras y segmentos del corpus de entrenamiento original.

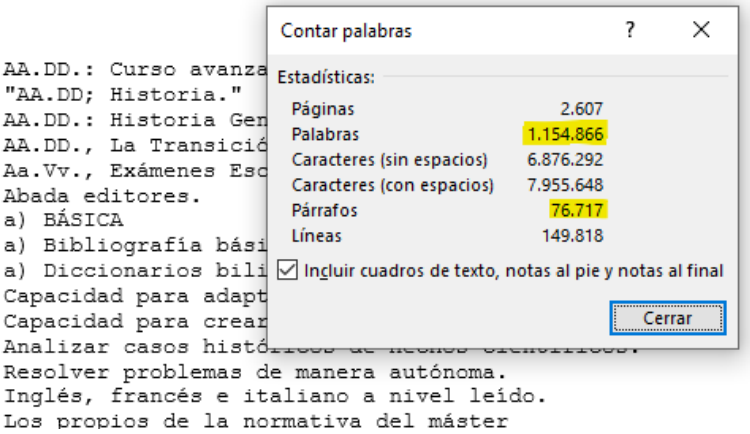


Figura 2. Conteo de palabras y segmentos del corpus de entrenamiento limpio.

El corpus de optimización, por su parte, consta de 677 segmentos seleccionados de guías docentes del mismo currículo, alineados y revisados para confirmar que su traducción sea correcta. Dichos segmentos contienen 7.659 palabras en español y 7.484 palabras en inglés. Durante el proceso de

limpieza, descrito en el apartado 5.1.3., el número de segmentos se redujo a 342. Es decir, 4.643 palabras en español y 4.611 en inglés.

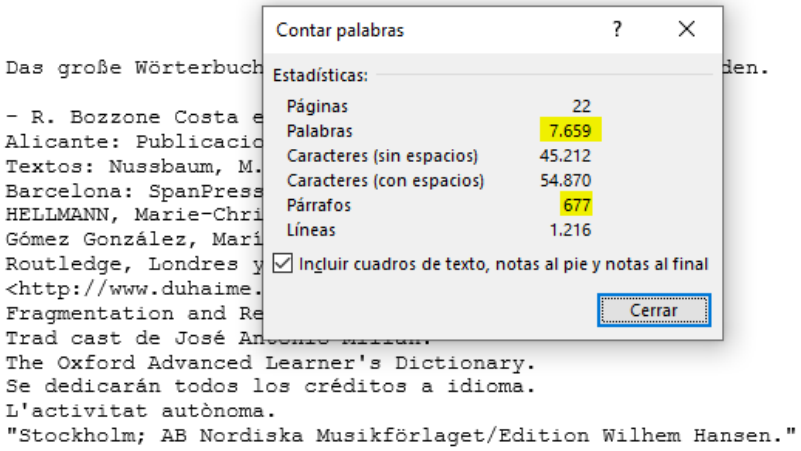


Figura 3. Conteo de palabras y segmentos del corpus de optimización original.

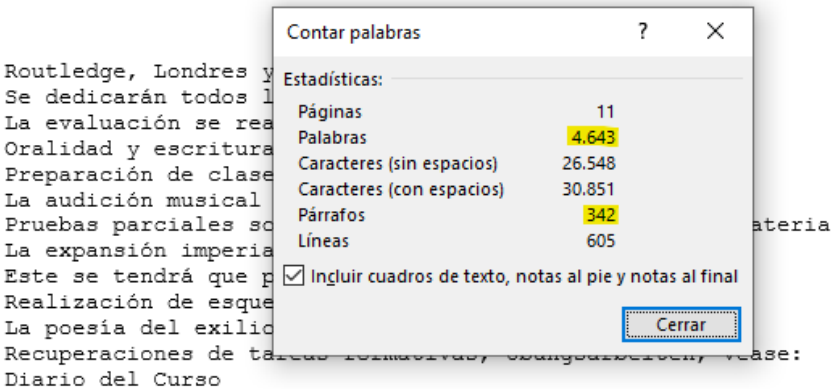


Figura 4. Conteo de palabras y segmentos del corpus de optimización limpio.

Los motores de entrenamiento se prueban con cinco guías docentes representativas del mismo currículo, pero que no fueron incluidas ni en el corpus de entrenamiento ni en el de optimización. Estas guías son relativamente cortas, y pertenecen a grados como historia, psicología y traducción.

5.1.2. Preparación del Corpus de Entrenamiento

El primer paso fue limpiar el corpus de entrenamiento, es decir, eliminar todos los segmentos que se encontraran en un idioma diferente al español o

el inglés, o aquellos que no ofrecieran ninguna información relevante para efectos del entrenamiento del motor de traducción automática, tales como los que sólo contuvieran nombres propios, marcas comerciales o números.

EN	ES
Aachen: Shaker.	Aachen: Shaker.
Aachen: Shaker Verlag.	Aachen: Shaker Verlag.
AA.DD.: Curso avanzado de italiano.	AA.DD.: Curso avanzado de italiano.
AA.DD., Diccionari de la llengua catalana, Barcelona:	AA.DD., Diccionari de la llengua catalana, Barcelona:
AA.DD, Diccionario general de la lengua española,.	AA.DD, Diccionario general de la lengua española,.
AA.DD; Història.	AA.DD; Història.
AA.DD.: Historia General de Africa.	AA.DD.: Historia General de Africa.
AA.DD., La Transición, treinta años después.	AA.DD., La Transición, treinta años después.

Figura 5. Segmentos a limpiar del corpus de entrenamiento.

Una vez estaba limpio el corpus, se procedió a realizar la traducción automática del mismo al idioma de destino (en este caso el inglés) por medio del motor Google Translate. Se escogió trabajar con este motor ya que, a diferencia de otros motores de TA tales como DeepL, Google Translate permite su uso para fines de investigación y de evaluación de motores de TA. Acto seguido, se hizo una posedición ligera de la traducción resultante. Se cambiaron inconsistencias terminológicas entre segmentos, y falsos amigos.

Reconstructing the artistic panorama of the contemporary wo	Reconstructing the artistic panorama of the contemporary world.
Registration methodologies.	Registration methodologies.
-Records: Descriptive Units.	-Records: Descriptive Units.
RECOVERY: You can only recover written work.	makeup exam: You can only recover written work.
Recovery process	makeup exam process
Recovery consist of a final exam synthesis.	makeup exam consist of a final exam synthesis.
Recovery:	makeup exam:
Recovery	makeup exam
RECOVERY	makeup exam
Morales recovery of some concepts of Ancient Greece.	Morales makeup exam of some concepts of Ancient Greece.
Information Resources (Library Service UAB):	Information Resources (Library Service UAB):

Figura 6. Posedición del corpus de entrenamiento.

5.1.3. Preparación del corpus de optimización

Al igual que con el corpus de entrenamiento, el primer proceso que se realizó fue la limpieza del corpus de optimización, con las mismas especificaciones. Es decir, se eliminaron los segmentos en otros idiomas o que contuvieran datos irrelevantes para efectos de la traducción. Este corpus limpio se guardó en dos versiones diferentes.

El siguiente paso fue realizar una preedición del mismo, ya que presentaba algunos problemas que afectarían la calidad de la traducción, principalmente de espaciado, puntuación, y presencia de palabras escritas en catalán en textos en español.

Después, se tradujeron todos los segmentos por medio del motor de traducción de Google Translate y, por último, una posedición ligera de la traducción resultante.

Se probaron las dos versiones del corpus de optimización limpio, y debido a la consistencia de resultados, se decidió tomar como corpus definitivo para este trabajo el que sólo había sido sometido al proceso de limpieza, pero no al de preedición, traducción y posedición.

5.2. Implementación del instrumento

5.2.1. Entrenamiento de MTradumática

Se crearon cuatro motores de traducción: uno basado en el corpus de entrenamiento original (es decir, tal como se encontraba antes de pasar por el proceso de limpieza, preedición, y posedición), y el otro basado en el corpus de entrenamiento limpio, después de pasar por dicho proceso. Asimismo, cada uno de estos motores fue creado en dos versiones según el corpus de optimización utilizado: el corpus original, y el corpus revisado y poseditado.

Después se entró en la plataforma de MTradumática para realizar el proceso de entrenamiento de los motores. Para ello, se siguieron los siguientes pasos, tal como se delinean en la página de la plataforma:

Cargar los ficheros: Estos son los archivos en formato .txt de los corpus de entrenamiento y de optimización. Cada corpus en el idioma de partida (en

este caso, español), va en un archivo de texto, y cada corpus en el idioma de llegada, es decir en inglés, va en un archivo de texto por separado.

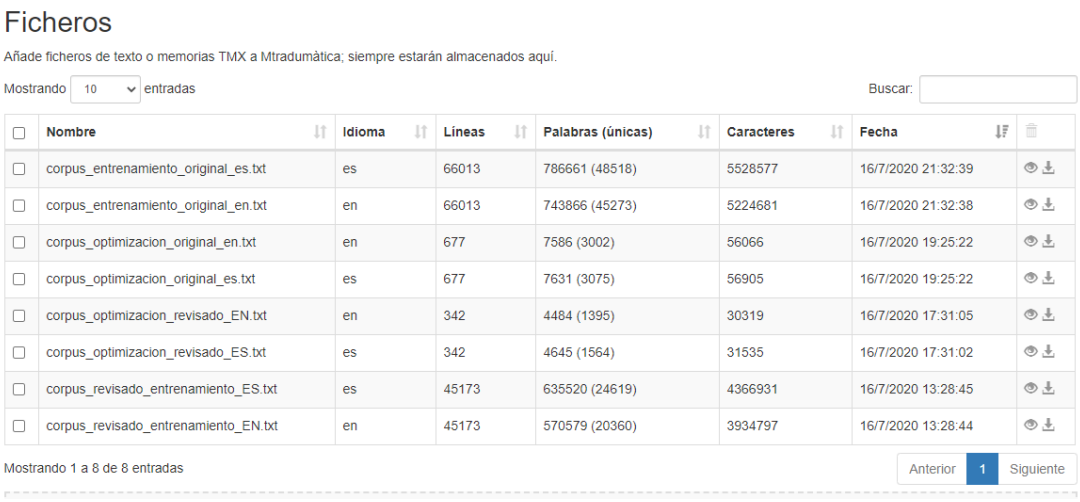


Figura 7. Cargando los ficheros a la plataforma de MTradumática.

Crear monotextos: Estos son creados a partir de los ficheros cargados en la plataforma.

Figura 8. Creación de monotextos.

Entrenar el modelo de lengua: a partir de los monotextos en la lengua de llegada.

Figura 9. Entrenamiento de los modelos de lengua.

Crear bitextos: En este paso cada monotexto en el español se reúne con su correspondiente monotexto en inglés, para crear el bitexto que será la base del motor de TA.

Figura 10. Creación de bitextos.

Crear y entrenar los motores de traducción: Cada motor se crea y entrena a partir de un bitexto del corpus de entrenamiento, ya sea el original o el limpio. La plataforma tarda unos cuantos minutos en entrenar cada motor.

Figura 11. Entrenamiento de traductores automáticos.

Optimizar los motores de traducción: Una vez creado y entrenado el motor de traducción automática, se puede llevar a cabo de manera opcional el proceso de optimización, a partir de un bitexto del corpus de optimización original o limpio según sea el caso. Este proceso suele llevar mucho más tiempo que el de entrenamiento.

Figura 12. Tiempos de entrenamiento y tiempos de optimización.

En total se crearon seis motores de traducción automática a partir de las combinaciones entre corpus de entrenamiento y corpus de optimización:

- corpus de entrenamiento original solo
- corpus de entrenamiento limpio
- corpus de entrenamiento original + corpus de optimización original
- corpus de entrenamiento original + corpus de optimización limpio
- corpus de entrenamiento limpio + corpus de entrenamiento original
- corpus de entrenamiento limpio + corpus de entrenamiento limpio

5.2.2. Preparación de las guías docentes

Una vez los motores de traducción automática estuvieron listos, se procedió a preparar las cinco guías docentes para su traducción. Cada una de las guías, originalmente en español, venía con su traducción al inglés, realizada por los

mismos docentes que las crearon. La preparación consistió principalmente en alinear los segmentos del idioma de partida con los de llegada, y en realizar una posedición ligera de las traducciones al inglés de cada una de las guías. Cada guía se guardó en dos versiones: la versión original, y la versión poseditada. Cada versión se guardó en dos archivos en formato .txt, separando la guía en español de su correspondiente traducción al inglés.

5.2.3. Traducción de las guías docentes

Al estar las guías preparadas, el siguiente paso fue traducir cada una las cinco guías docentes en español, en cada una de sus dos versiones, usando cada uno de los seis motores de traducción automática. En total se hicieron 60 traducciones automáticas.

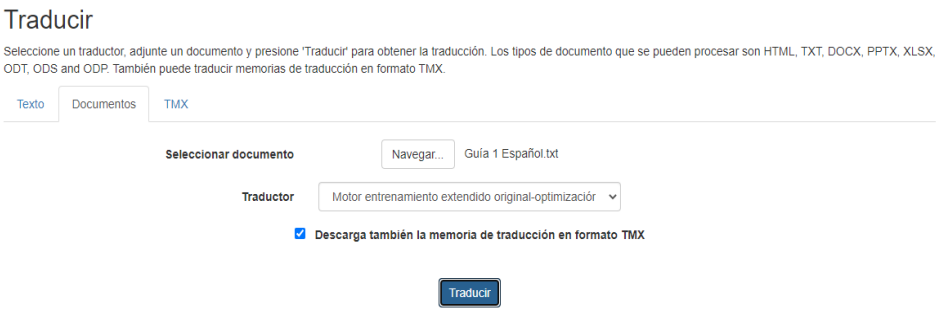


Figura 13. Traducción de documentos con MTradumática.

Adicionalmente, se hizo el mismo proceso usando el motor de traducción automática de Google Translate, creando otras 10 traducciones para evaluar y comparar este motor con los creados por la plataforma de MTradumática.

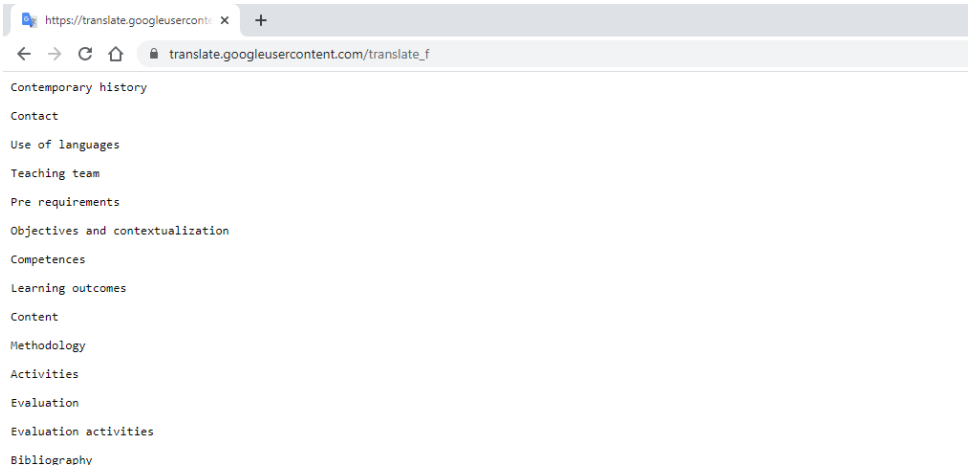


Figura 14. Traducción de documentos con Google Translate.

5.2.4. Comparación De Motores De Traducción

Para realizar la evaluación de los diferentes motores de traducción, se utilizó la métrica BLEU, descrita en el apartado 4.4.1., a través de la plataforma Tilde.

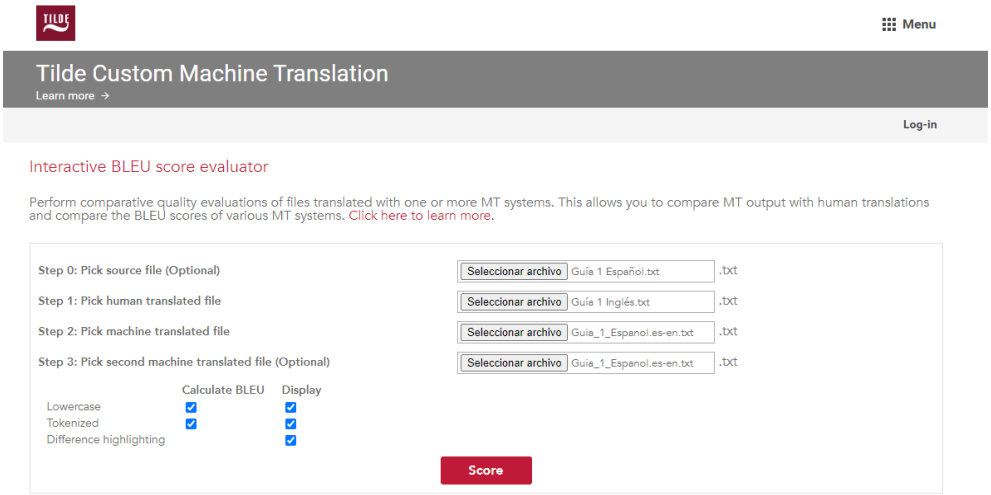


Figura 15. Evaluador BLEU de la plataforma Tilde.

Para realizar la evaluación, se subieron a la plataforma: el archivo original de cada guía en la lengua de partida (español), el archivo que contenía la traducción humana de dicha guía, con la que se comparan los motores; y dos de las traducciones automáticas de la misma cada vez, ya que es el número máximo que la plataforma permite comparar al mismo tiempo. De esta

manera, cada una de las 70 traducciones se evaluó siguiendo el mismo procedimiento.

Al presentar la evaluación, la plataforma Tilde muestra el puntaje BLEU de cada una de las traducciones automáticas expresándolas en porcentajes, es decir, de 0 a 100, lo cual permite una interpretación más fácil de los resultados. Igualmente muestra el proceso de evaluación tanto individual como acumulado, desde el 1-grama hasta el 4-grama, siendo el puntaje acumulado 4-grama el que se considera como puntaje global. Esto significa que según la métrica BLEU, la calificación se basa en la probabilidad de aparición de una secuencia de 4 palabras en cada segmento, similar a la de la traducción humana con la que se compara.

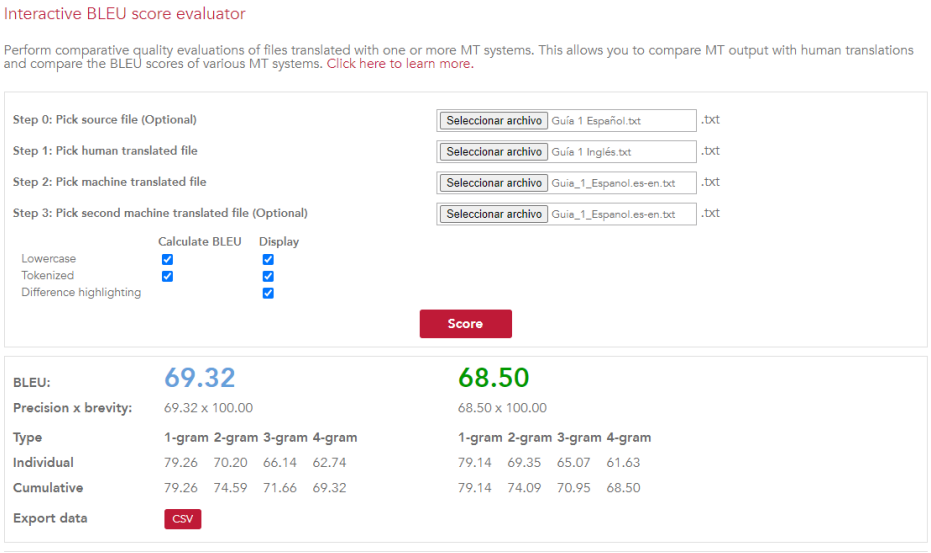


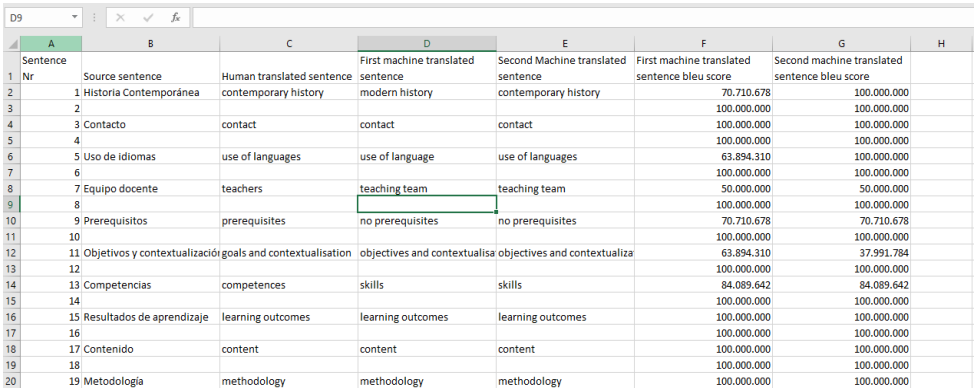
Figura 16. Puntaje global BLEU.

También muestra el puntaje BLEU desglosado segmento a segmento, tanto en forma de gráfico como en forma de tabla, detallando las diferencias entre las traducciones de cada motor.

Sentence 1	BLEU	Length ratio	Text
Source	-	-	Historia Contemporánea
Human	100.00	1.00	Contemporary History
Machine	70.71	1.00	Modern History
Machine	70.71	1.00	Modern History

Figura 17. Puntaje BLEU desglosado por segmentos.

Aunque el gráfico y el puntaje global no se pueden descargar, sí se puede descargar el archivo .csv de la comparación de los puntajes BLEU segmento a segmento.



Sentence Nr	Source sentence	Human translated sentence	First machine translated sentence	Second Machine translated sentence	First machine translated sentence bleu score	Second machine translated sentence bleu score
1	Historia Contemporánea	contemporary history	modern history	contemporary history	70.710.678	100.000.000
2					100.000.000	100.000.000
3	Contacto	contact	contact	contact	100.000.000	100.000.000
4					100.000.000	100.000.000
5	Uso de idiomas	use of languages	use of language	use of languages	63.894.310	100.000.000
6					100.000.000	100.000.000
7	Equipo docente	teachers	teaching team	teaching team	50.000.000	50.000.000
8					100.000.000	100.000.000
9	Prerequisitos	prerequisites	no prerequisites	no prerequisites	70.710.678	70.710.678
10					100.000.000	100.000.000
11	Objetivos y contextualizació	goals and contextualisation	objectives and contextualisa	objectives and contextualiza	63.894.310	37.991.784
12					100.000.000	100.000.000
13	Competencias	competences	skills	skills	84.089.642	84.089.642
14					100.000.000	100.000.000
15	Resultados de aprendizaje	learning outcomes	learning outcomes	learning outcomes	100.000.000	100.000.000
16					100.000.000	100.000.000
17	Contenido	content	content	content	100.000.000	100.000.000
18					100.000.000	100.000.000
19	Metodología	methodology	methodology	methodology	100.000.000	100.000.000

Figura 18. Archivo .csv descargable.

5.2.5. Verificación De Resultados

Después de realizar el anterior proceso, se llegó a la conclusión de que era necesario verificar los resultados obtenidos repitiendo el mismo con guías diferentes, sacadas de materias de dos programas de pregrado nuevos: Estudios de Español y Chino: Lengua, Literatura y Cultura, y Filología Catalana: Estudios de Literatura y Lingüística.

A diferencia de las guías originales, estas nuevas guías, que llamaremos guías complementarias, no contaban con una traducción humana. Por esta razón, el primer paso fue realizar las traducciones humanas de dichas guías utilizando la herramienta Memsource. Para evitar cualquier confusión a la hora de evaluar los motores, se evitó el uso de Google Translate o cualquier otra herramienta de traducción automática como ayudas en el proceso de traducción.

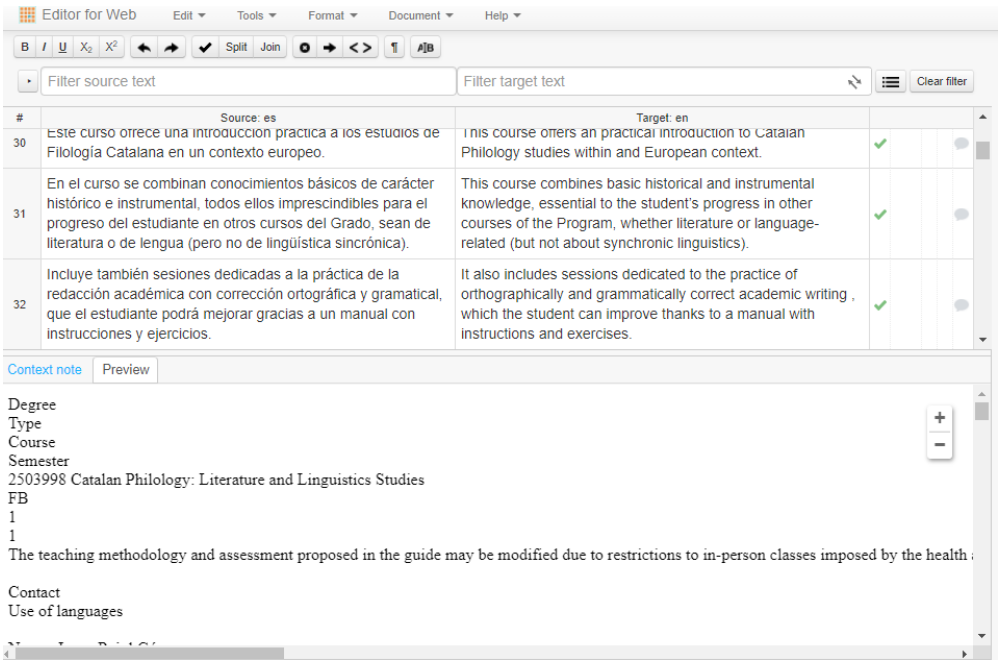


Figura 19. Traducción humana de las guías complementarias.

Acto seguido, se siguió exactamente la misma secuencia de pasos que con las guías originales.

6. ANÁLISIS DE LOS RESULTADOS

En esta sección veremos el análisis de los resultados obtenidos a través de la evaluación realizada según la métrica BLEU, tanto de las guías originales como de las guías complementarias, con el objetivo de definir cuál es el motor más apropiado para traducir las guías docentes de la UAB.

En el apartado 6.1. se encuentran las tablas en las que se organizan los puntajes BLEU de cada una de las guías según el motor utilizado. En el 6.2. se comparan los puntajes del motor de MTradumática con los de Google Translate, en el 6.3. se contrastan los motores que usan el corpus de entrenamiento original con los que usan el corpus de entrenamiento limpio y finalmente, en el apartado 6.4. se examina la diferencia entre los puntajes de los motores según el corpus de optimización que usan. Por último, en el apartado 6.5. se hará una breve comparación a simple vista entre traducción humana, la hecha por Google Translate, y la hecha por un motor de MTradumática.

6.1. Resultados generales

En las siguientes tablas se puede observar la comparación entre los diferentes puntajes BLEU de las traducciones de cada guía:

Guía 1

Guía 1		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	71,69	42,87
	Original	20,04	29,63
	Limpio	70,14	42,34
Google Translate		42,63	

Tabla 2. Resultados de la guía 1.

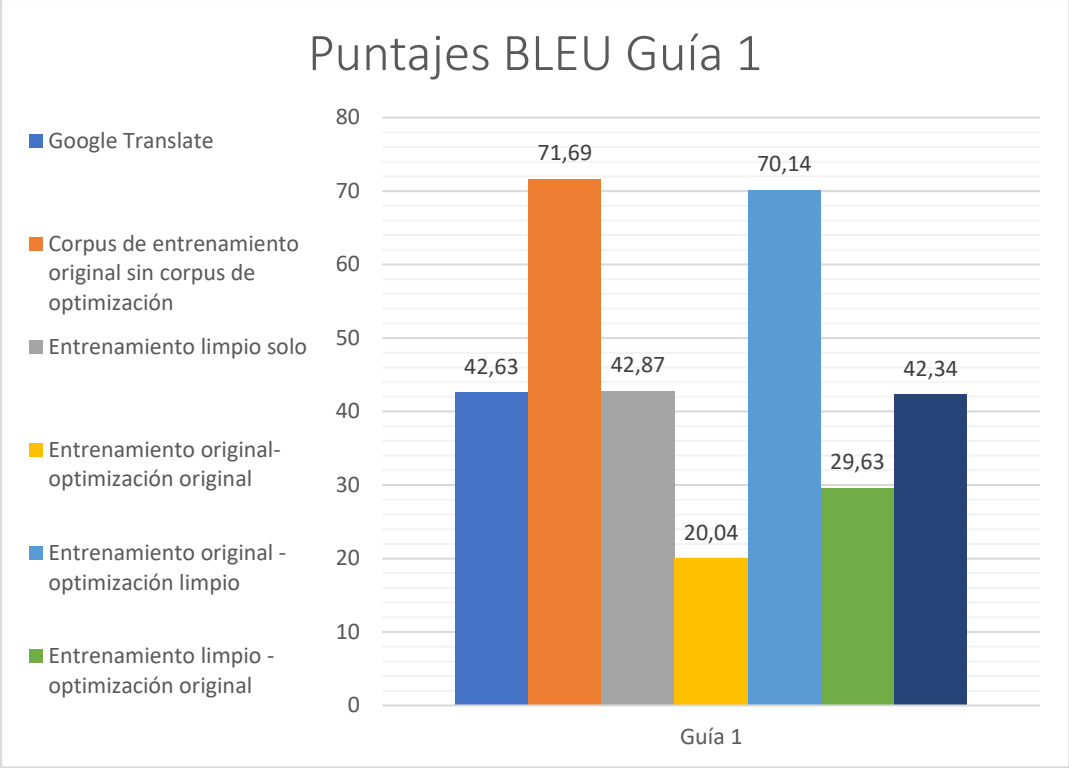


Figura 20. Comparación de puntajes BLEU de la guía 1.

Guía 2

Guía 2		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	50,5	38,7
	Original	23,65	24,24
	Limpio	45,37	37,17
Google Translate		42,08	

Tabla 3. Resultados de la guía 2.

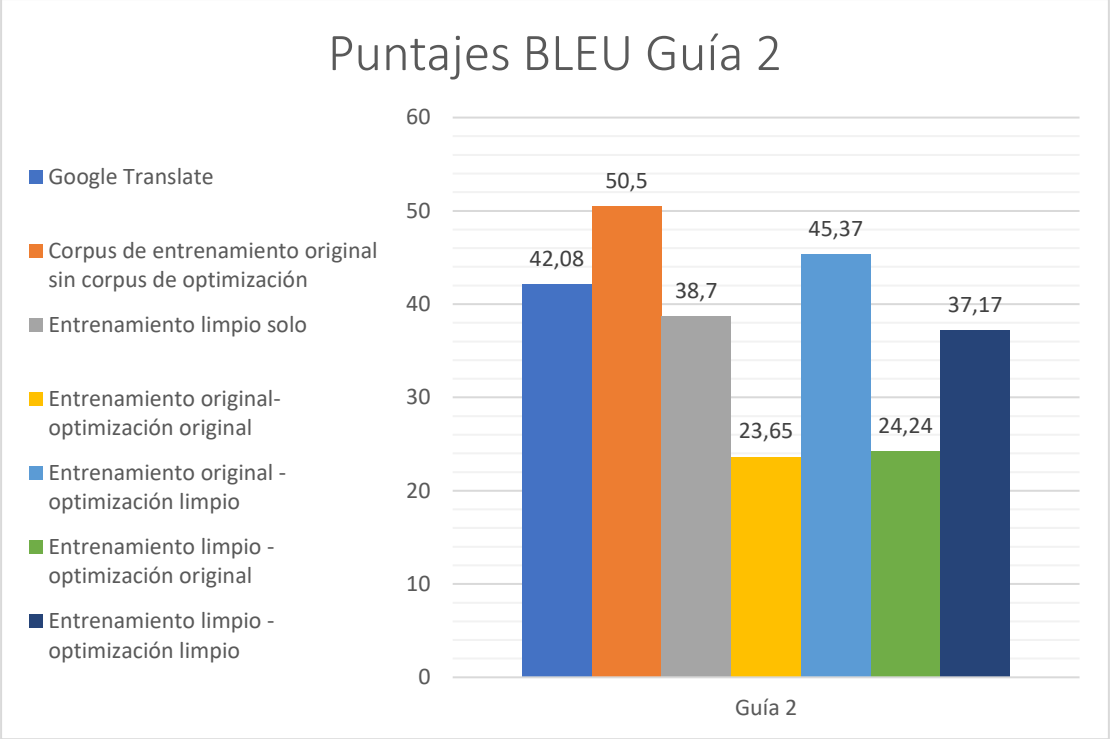


Figura 21. Comparación de puntajes BLEU de la guía 2.

Guía 3

Guía 3		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	54,38	48,23
	Original	10,83	28,68
	Limpio	56,03	49,98
Google Translate		49,9	

Tabla 4. Resultados de la guía 3.

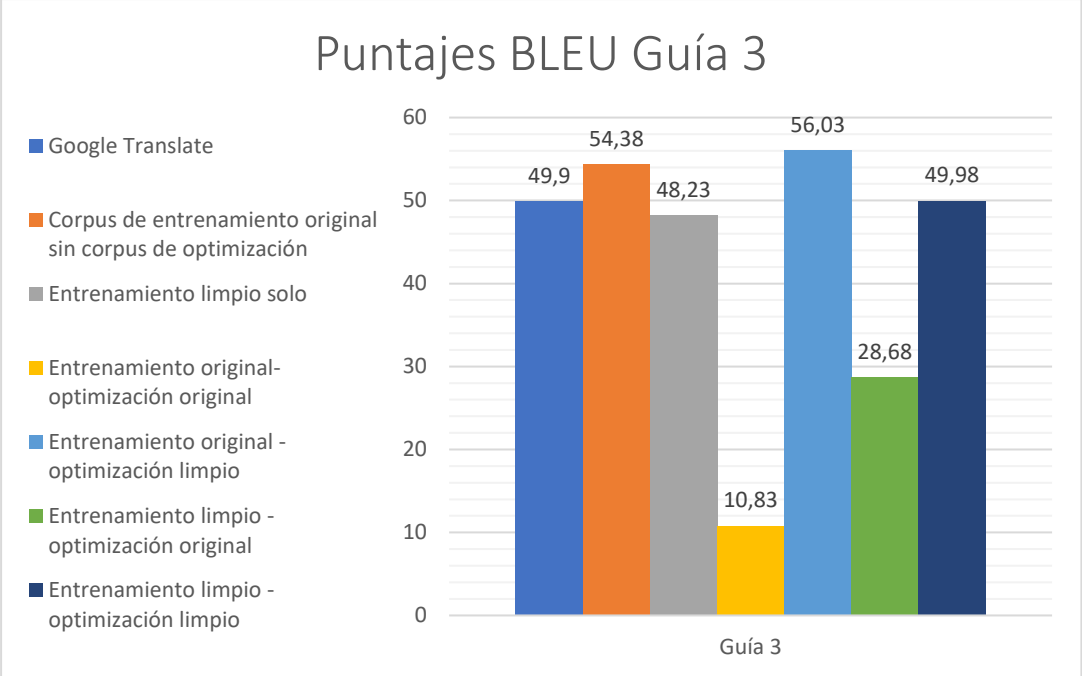


Figura 22. Comparación de puntajes BLEU de la guía 3.

Este es el único caso en el que uno de los motores, el entrenado con corpus de entrenamiento y optimización originales, obtuvo una puntuación por debajo de la mínima para poder captar la esencia de la traducción: 10,83. Esta se

Guía 4

Guía 4		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	70,54	49,29
	Original	25,28	30
	Limpio	69,93	50,66
Google Translate		51,19	

Tabla 5. Resultados de la guía 4.

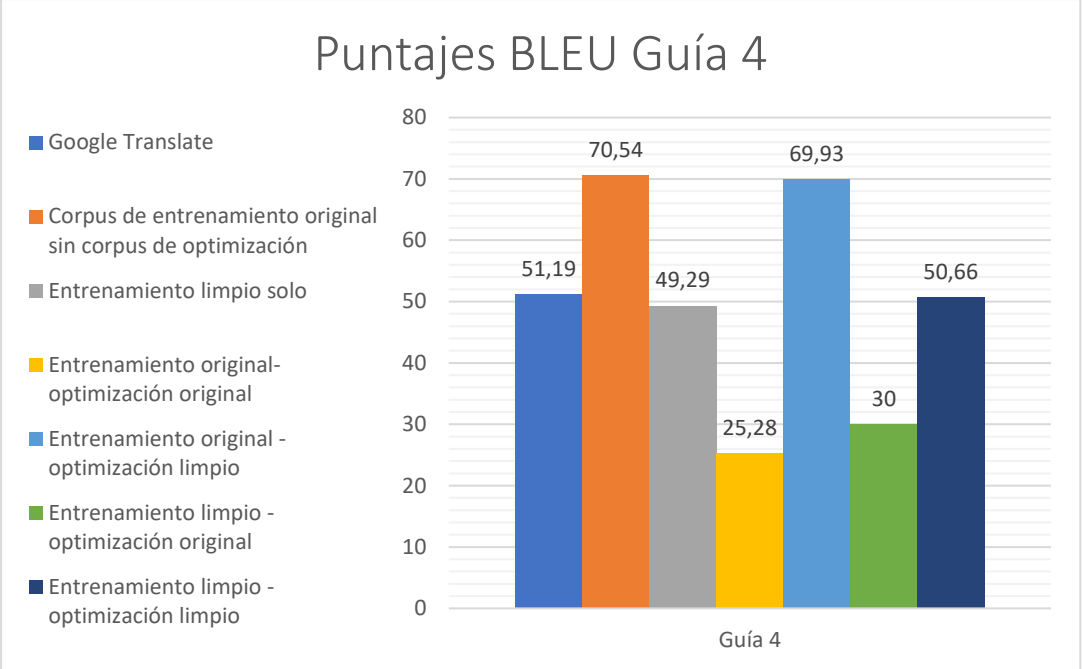


Figura 23. Comparación de puntajes BLEU de la guía 4.

Guía 5

Guía 5		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	80,25	50,64
	Original	27,95	35,6
	Limpio	81,37	55,57
Google Translate		50,27	

Tabla 6. Resultados de la guía 5.

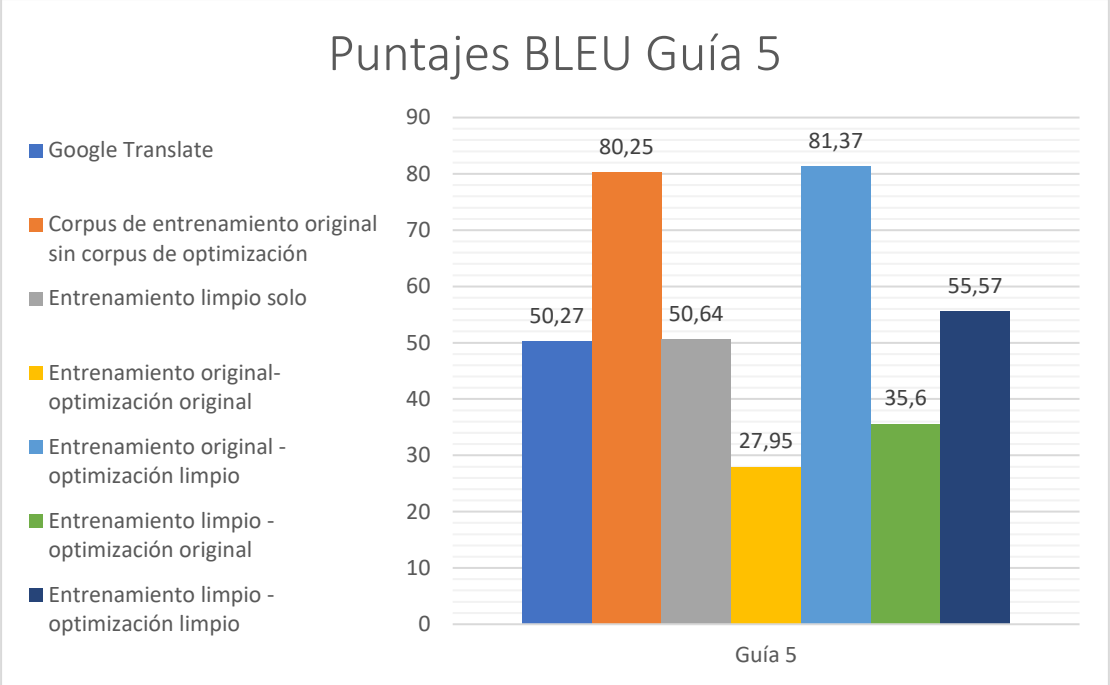


Figura 24. Comparación de puntajes BLEU de la guía 5.

Promedio

Promedio		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	65,47	45,95
	Original	21,55	29,63
	Limpio	64,57	47,14
Google Translate		47,21	

Tabla 7. Promedio de los resultados de las guías.

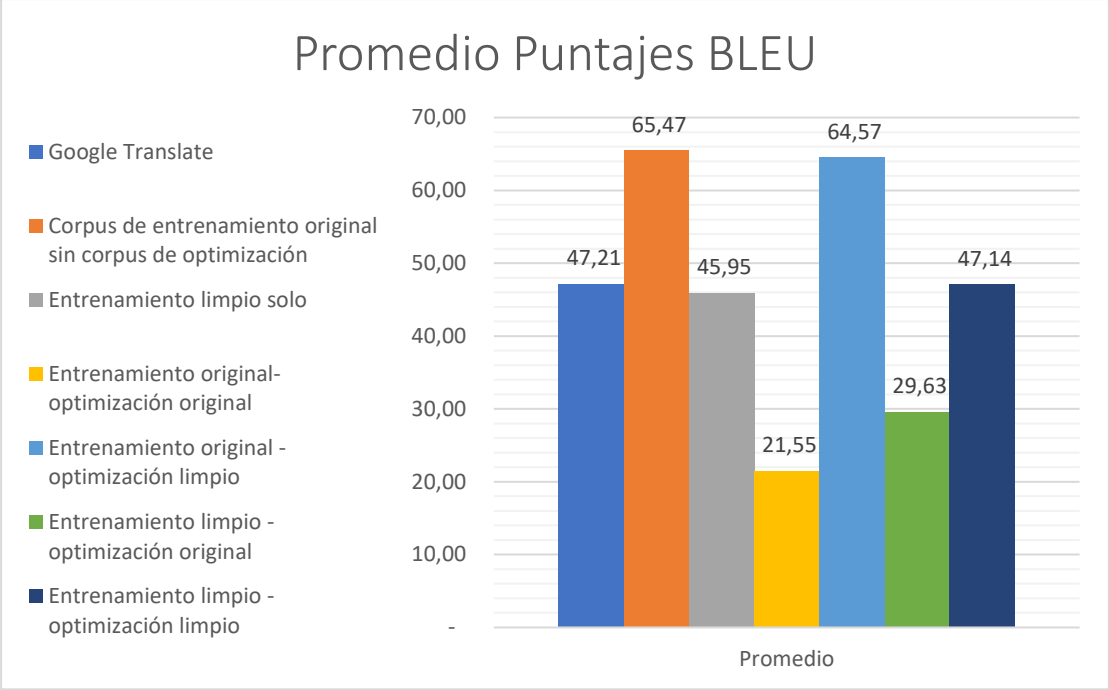


Figura 25. Comparación de promedios de puntajes BLEU de todas las guías.

Tal y como se puede observar en la figura 25, los puntajes obtenidos por los siete motores evaluados se pueden dividir en tres niveles bastante diferentes, según la tabla de puntajes de BLEU:

Esencia clara, pero con errores gramaticales significativos: En este nivel se encuentran el motor entrenado con el corpus de entrenamiento original y el corpus de optimización original (puntaje de 21,55), y el entrenado con el corpus de entrenamiento limpio y el corpus de optimización original (puntaje de 29,63).

Traducciones de alta calidad: Aquí se ubica el motor de Google Translate, con un puntaje de 47,21; así como el motor entrenado sólo con el corpus de entrenamiento limpio (45,95 puntos) y el entrenado con el corpus de entrenamiento y el de optimización limpios.

Calidad que suele ser mejor que la humana: En el nivel más alto se encuentran dos motores de MTradumática: el entrenado únicamente con el corpus de entrenamiento original, y el entrenado con el corpus de entrenamiento original y el corpus de optimización limpio.

Guías complementarias

Chino I

Chino 1		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	45,48	45,5
	Original	22,75	31,06
	Limpio	45,48	43,57
Google Translate		58,97	

Tabla 8. Resultados de la guía Chino 1.

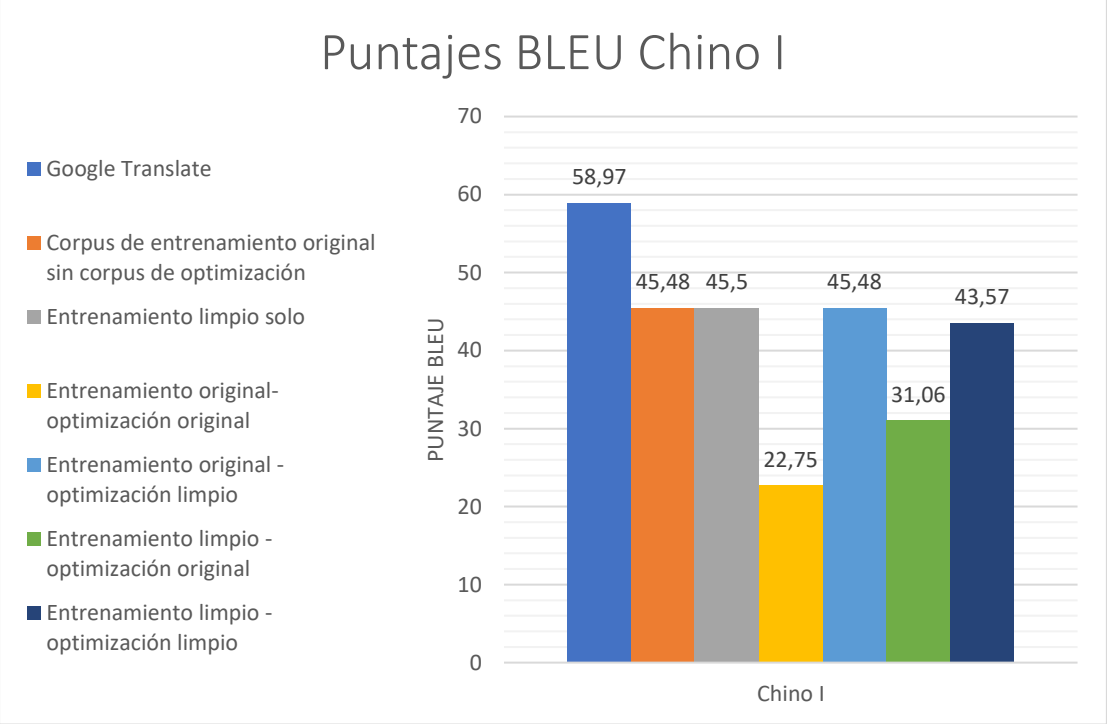


Figura 26. Comparación de puntajes BLEU de la guía Chino 1.

Chino II:

Chino 2		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	42,94	43,7
	Original	20,05	30,09
	Limpio	43,02	42,4
Google Translate		56,6	

Tabla 9. Resultados de la guía Chino 2.

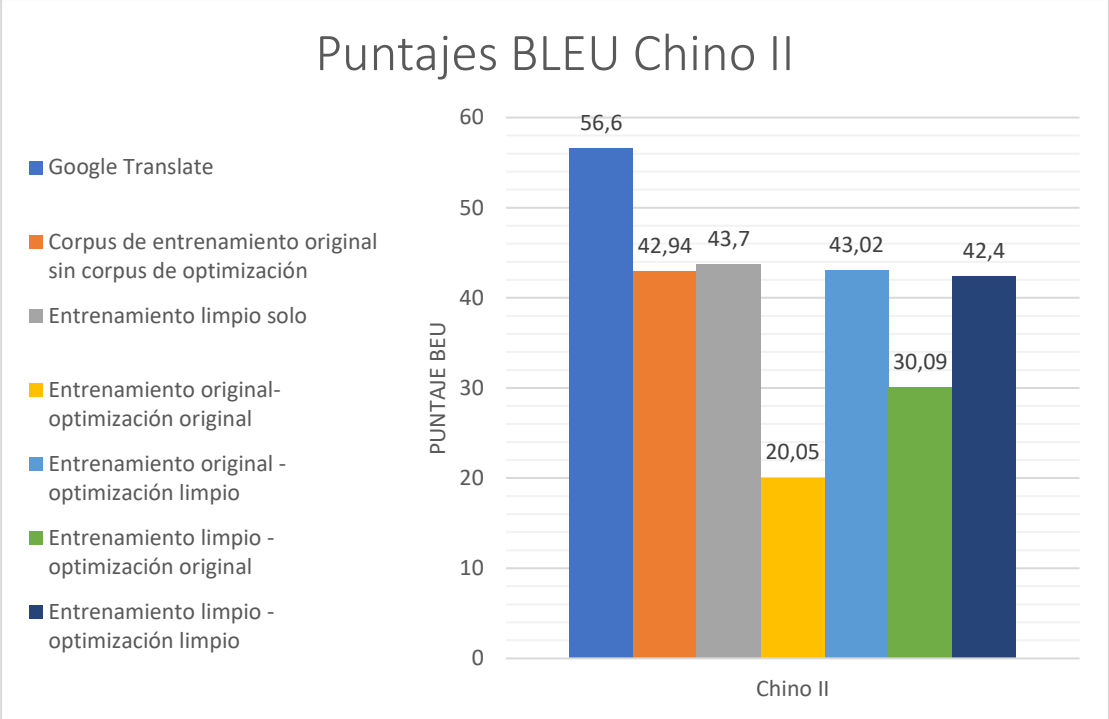


Figura 27. Comparación de puntajes BLEU de la guía Chino 2.

Filología Catalana:

Filología Catalana		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	55,19	48,98
	Original	25,2	37,28
	Limpio	55,46	48,04
Google Translate		59,61	

Tabla 10. Resultados de la guía de Filología Catalana.

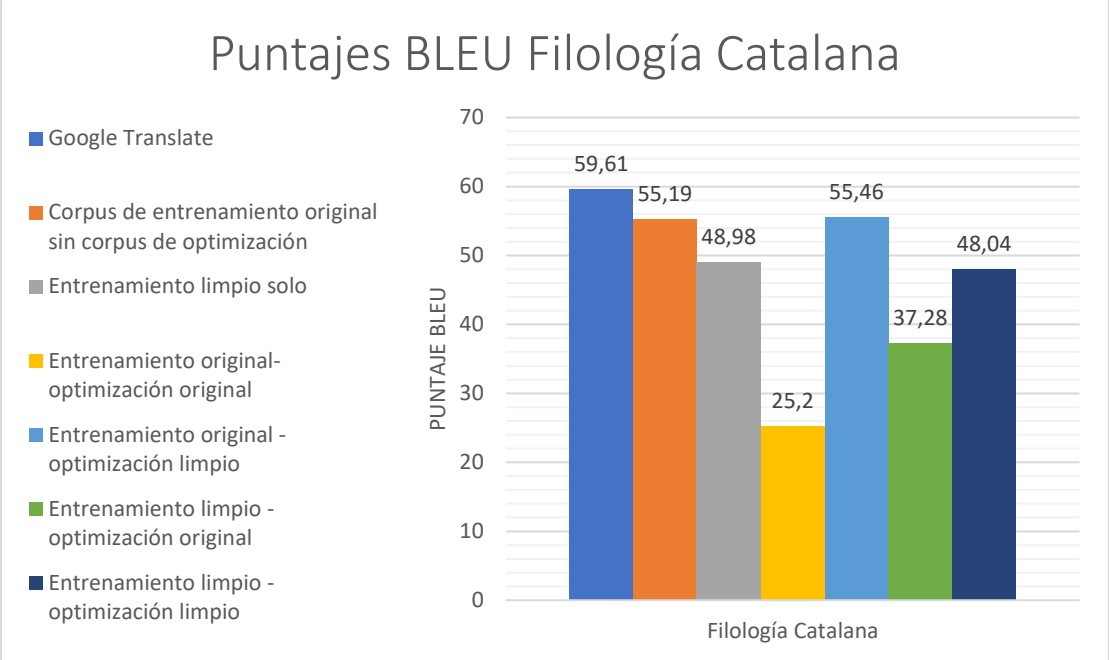


Figura 28. Comparación de puntajes BLEU de la guía Filología Catalana.

Promedio:

Promedio Guías Complementarias		Corpus entrenamiento	
		Original	Limpio
Corpus optimización	Ninguno	47,87	46,06
	Original	22,67	32,81
	Limpio	47,99	44,67
Google Translate		58,39	

Tabla 11. Promedio de resultados de las guías complementarias.

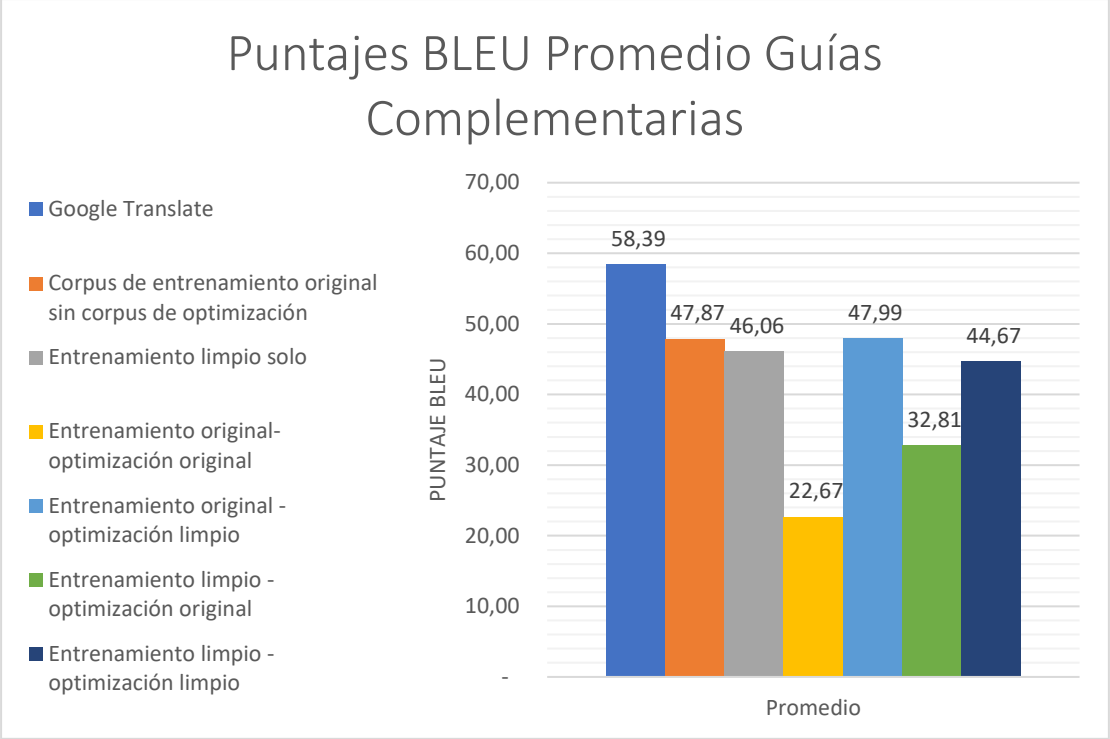


Figura 29. Comparación de promedios de puntajes BLEU de todas las guías complementarias.

La figura 29 nos muestra, al igual que en el caso de las guías originales, los puntajes obtenidos por los siete motores evaluados, que según la tabla de puntajes de BLEU, se dividen en cuatro niveles:

Esencia clara, pero con errores gramaticales significativos: En este nivel se encuentra el motor entrenado con el corpus de entrenamiento original y el corpus de optimización original (puntaje de 22,67).

En el nivel comprensible (puntaje entre 30 y 40) se ubica el motor entrenado con el corpus de entrenamiento limpio y el corpus de optimización original (puntaje de 32,81).

Traducciones de alta calidad: Aquí se ubican cuatro motores: el entrenado sólo con el corpus de entrenamiento original, con un puntaje de 47,87; así como el motor entrenado sólo con el corpus de entrenamiento limpio (46,06 puntos) el entrenado con corpus de entrenamiento original y corpus de optimización limpio (47,99 puntos),

y el entrenado con el corpus de entrenamiento y el de optimización limpios (44,67 puntos).

Traducciones de muy alta calidad: Aquí se ubica el motor de Google Translate, con un puntaje de 58,39.

A diferencia de las guías originales, ninguna de las traducciones de las guías complementarias alcanza una calidad que suele ser mejor que la humana.

En los próximos apartados se desglosarán los resultados con el fin de dilucidar mejor qué aspectos influyen en la diferencia de calidad de las traducciones evaluadas.

6.2. MTradumática vs Google Translate

En todas las guías originales, las traducciones hechas con el motor de MTradumática entrenado con el corpus de entrenamiento original y las hechas con el motor entrenado con el corpus de entrenamiento original y el corpus de optimización limpio tienen un puntaje bastante superior al de las hechas en Google Translate.

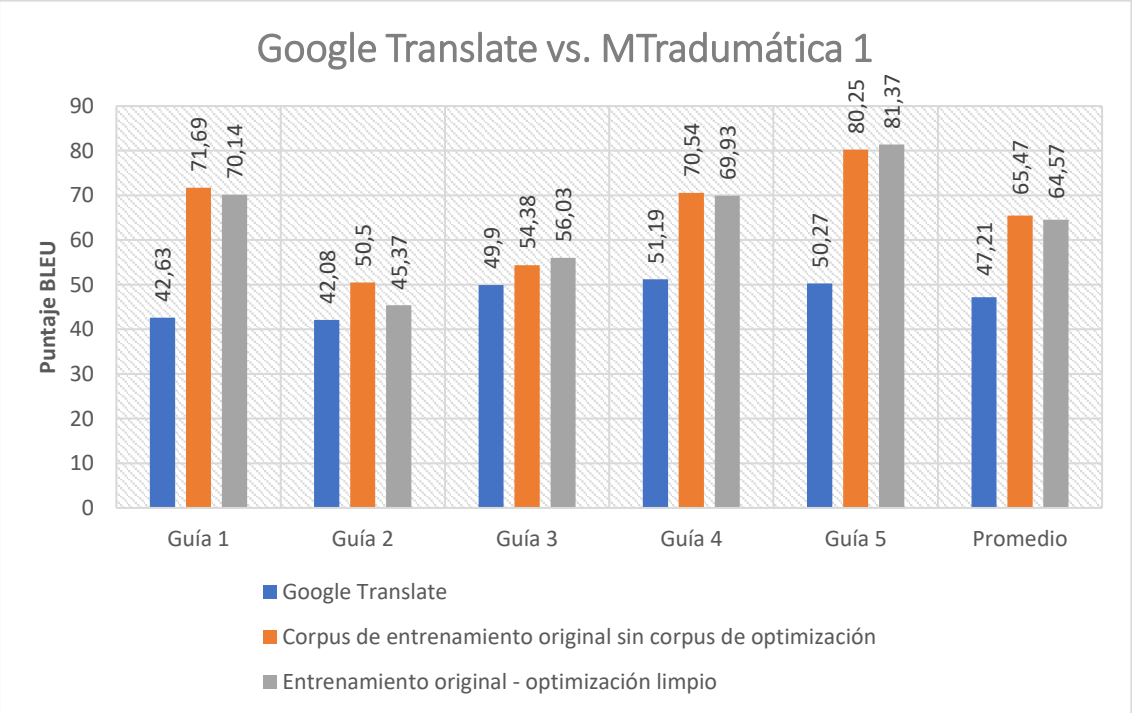


Figura 30. Puntajes BLEU de guías traducidas con Google Translate vs. motor de MTradumática entrenado con el corpus de entrenamiento original sin corpus de

optimización vs. el motor de MTradumática entrenado con el corpus de entrenamiento original y corpus de optimización limpio.

La diferencia promedio entre los puntajes BLEU de las traducciones hechas con Google Translate y las realizadas con el motor de MTradumática – corpus de entrenamiento original es de 18,26 puntos, La diferencia en este caso es de 17,36 puntos. Esta es una diferencia muy significativa a favor del motor de traducción automática especializado. Si bien según la escala, Google Translate produce traducciones calificada como de alta calidad, con un puntaje de 47,21, las traducciones producidas por el motor entrenado por MTradumática, con un puntaje promedio de 65,47, están dos niveles por encima, siendo calificadas como traducciones con calidad "que suele ser mejor que la humana".

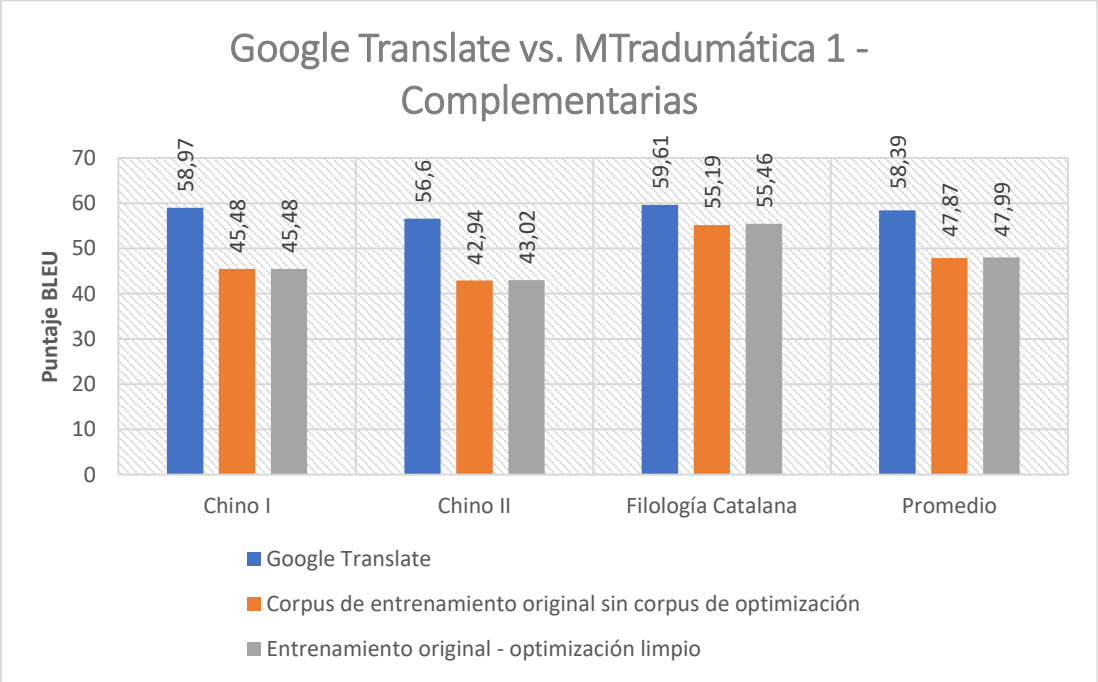


Figura 31. Puntajes BLEU de las guías complementarias traducidas con Google Translate vs. motor de MTradumática entrenado con el corpus de entrenamiento original sin corpus de optimización vs. el motor de MTradumática entrenado con el corpus de entrenamiento original y corpus de optimización limpio.

Por otro lado, en el caso de las guías complementarias, las traducciones hechas con Google Translate siempre superan a las realizadas con el motor de MTradumática. La diferencia entre el puntaje promedio alcanzado con Google Translate y el alcanzado

por el motor entrenado con el corpus de entrenamiento original sin corpus de optimización es de 10,52 puntos a favor del motor de Google Translate. La diferencia entre este último y el motor entrenado con corpus de entrenamiento original y corpus de optimización limpio es de 10,4 puntos. En estos casos la diferencia representa un salto de solamente una categoría: los motores entrenados por MTradumática.

Aquí se evidencia la mayor diferencia entre las guías originales y las complementarias: las traducciones de las originales llegan a puntajes considerados como comparables o superiores a las traducciones humanas, mientras que, en el caso de las guías complementarias, las mismas no alcanzan un puntaje de 50 sobre 100. Es decir, de acuerdo con el sistema BLEU, incluso las mejores traducciones hechas por MTradumática necesitan un porcentaje importante de edición humana. Aunque no se puede presumir a ciencia cierta la causa de esta discrepancia, cabe la posibilidad de que esto se deba a que las traducciones humanas de las guías originales fueron realizadas por los mismos profesores, al igual que el contenido de los corpus de entrenamiento y optimización, las traducciones de las guías complementarias fueron hechas por un tercero (en este caso, yo).

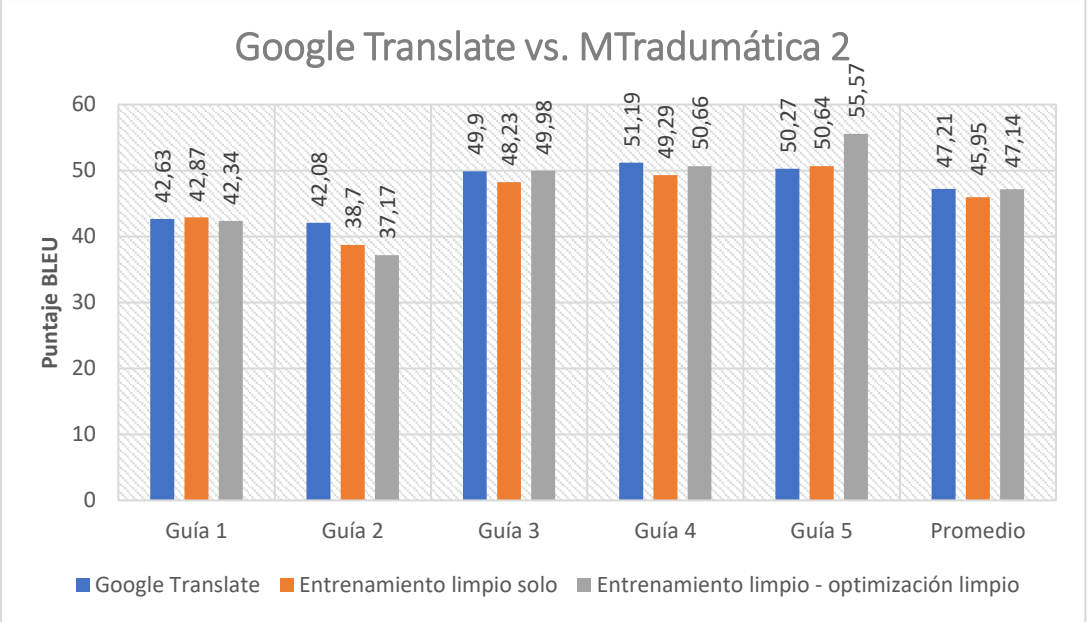


Figura 32. Puntajes BLEU de guías traducidas con Google Translate vs. motor de MTradumática entrenado con el corpus de entrenamiento limpio sin corpus de optimización vs. el motor de MTradumática entrenado con el corpus de entrenamiento y optimización limpios.

En este caso, la diferencia entre el puntaje de las traducciones hechas con Google Translate y las hechas con MTradumática tiende a ser insignificante, y los puntajes de los tres motores, con la excepción de los motores de MTradumática en la guía 2, rondan la mitad de la escala (un poco mas o menos de 50 sobre 100).

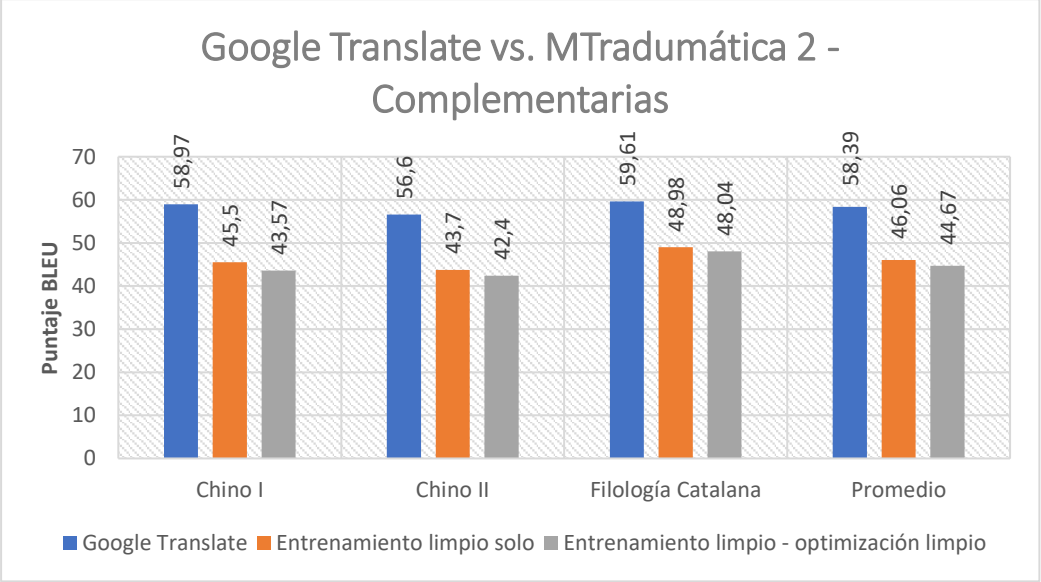


Figura 33. Puntajes BLEU de las guías complementarias traducidas con Google Translate vs. motor de MTradumática entrenado con el corpus de entrenamiento

limpio sin corpus de optimización vs. el motor de MTradumática entrenado con el corpus de entrenamiento y optimización limpios.

En el caso de las guías complementarias, además de la mayor diferencia entre los puntajes de las traducciones hechas con Google Translate y las hechas con los motores de MTradumática relacionados en la figura 32, que favorecen a las primeras, los puntajes de las últimas son parecidos a los de las traducciones de las guías originales.

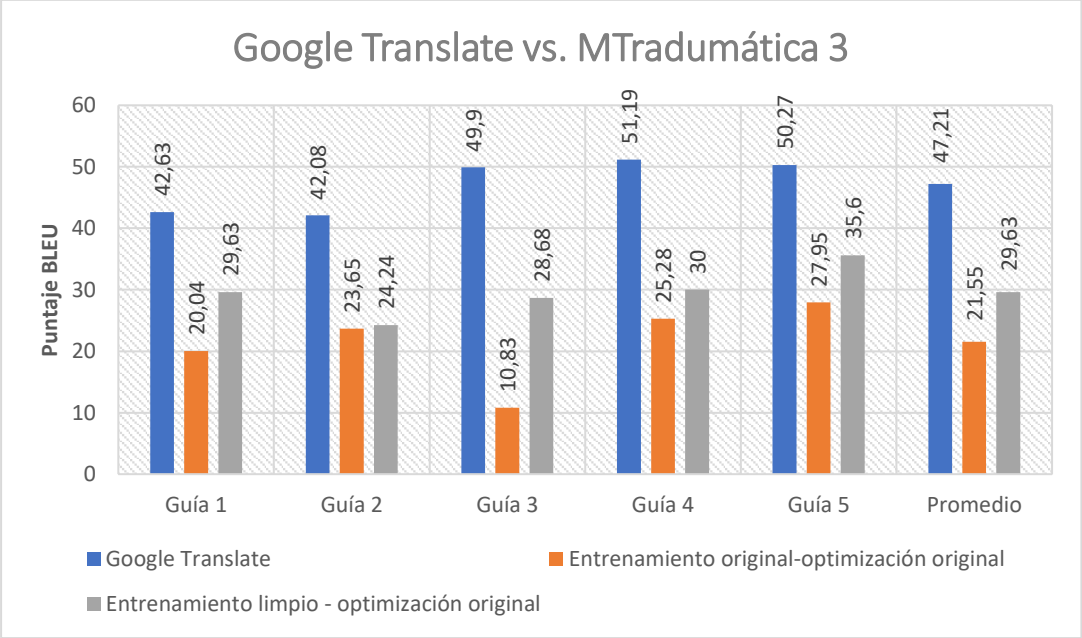


Figura 34. Puntajes BLEU de guías traducidas con Google Translate vs. motor de MTradumática entrenado con el corpus de entrenamiento y optimización originales vs. motor de MTradumática entrenado con el corpus de entrenamiento limpio y el corpus de optimización original.

En este caso, los puntajes de las traducciones hechas con los motores de MTradumática que aparecen en la figura 33 son bastante inferiores a los de las hechas con Google Translate. En la guía 3, el motor entrenado con corpus de entrenamiento y optimización originales sólo llega a un puntaje de 10,83. Esto significa que dicha traducción ni siquiera es usable.

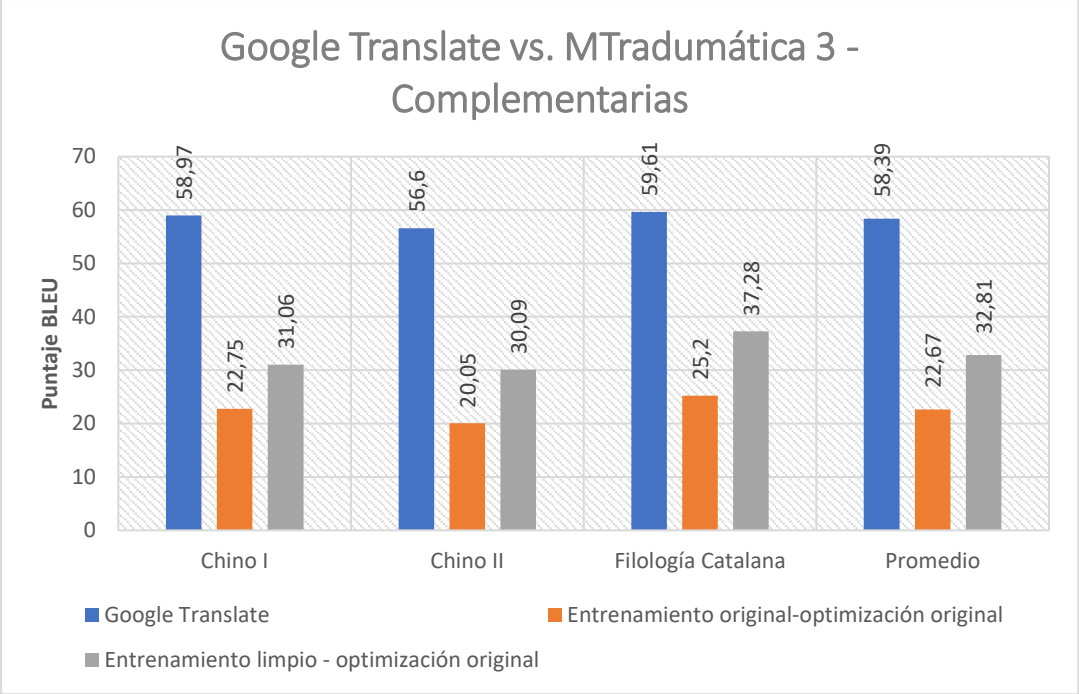


Figura 35. Puntajes BLEU de las guías complementarias traducidas con Google Translate vs. motor de MTradumática entrenado con el corpus de entrenamiento y optimización originales vs. motor de MTradumática entrenado con el corpus de entrenamiento limpio y el corpus de optimización original.

De nuevo, los puntajes de las traducciones hechas con los motores de MTradumática que aparecen en la figura 34 son bastante inferiores a los de las hechas con Google Translate, y muy similares (aunque con una variación menos amplia) a los de las traducciones de las guías originales entrenadas con estos mismos motores. Es decir, tanto el motor entrenado con corpus de entrenamiento limpio y corpus de optimización original como el entrenado con corpus de entrenamiento y de optimización originales se pueden descartar a la hora de escoger un motor para traducir las guías docentes de la UAB.

6.3. Corpus de entrenamiento original vs corpus de entrenamiento limpio

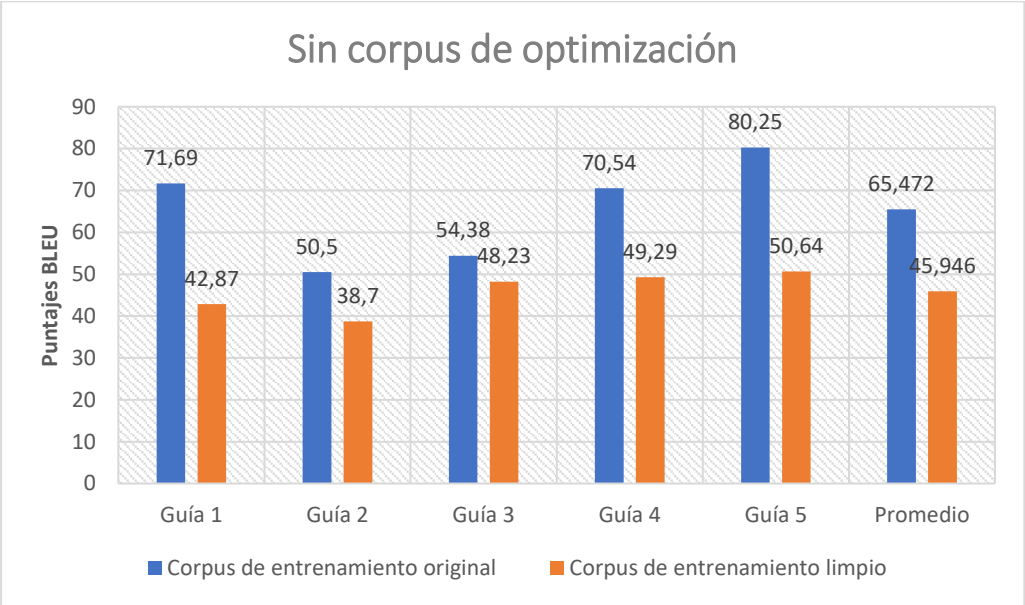


Figura 36. Puntajes BLEU de guías traducidas con el motor de MTradumática entrenado con el corpus de entrenamiento original vs. con el corpus de entrenamiento limpio, sin corpus de optimización.

En la figura 36 se observa un fenómeno interesante: mientras que las traducciones hechas solo con el corpus de entrenamiento limpio alcanzan puntajes medianos, las hechas únicamente con el corpus de entrenamiento original obtienen puntajes elevados, a tal punto que la traducción de la guía 5 obtiene un puntaje de 80,25, lo que es excepcionalmente alto y, al menos en el papel, potencialmente mejor que una traducción humana. Esto daría a entender que, en teoría, sería innecesario limpiar el corpus de entrenamiento.

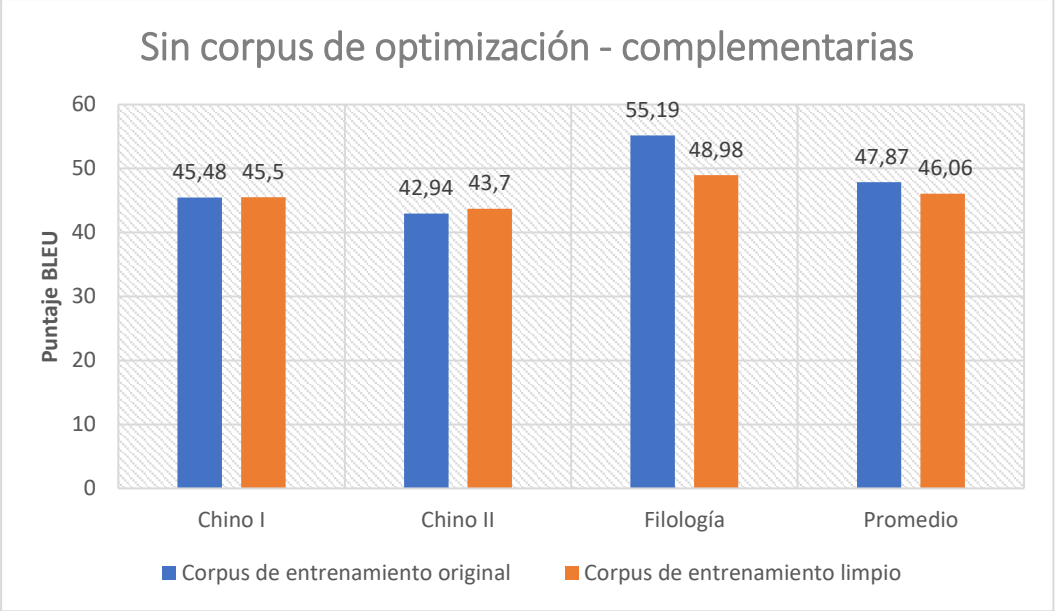


Figura 37. Puntajes BLEU de las guías complementarias traducidas con el motor de MTradumática entrenado con el corpus de entrenamiento original vs. con el corpus de entrenamiento limpio, sin corpus de optimización.

Sin embargo, al evaluar las traducciones de las guías complementarias, los puntajes del motor entrenado sólo con el corpus de entrenamiento original bajan considerablemente. Incluso en dos de las guías, el puntaje es levemente inferior al del motor entrenado solo con el corpus de entrenamiento limpio. Como se ha mencionado anteriormente, es probable que tal diferencia se deba al hecho de que los mismos profesores fueron quienes realizaron las traducciones de las guías originales, mientras que las guías complementarias fueron traducidas por mí, es decir, por un tercero con un estilo de traducción diferente.

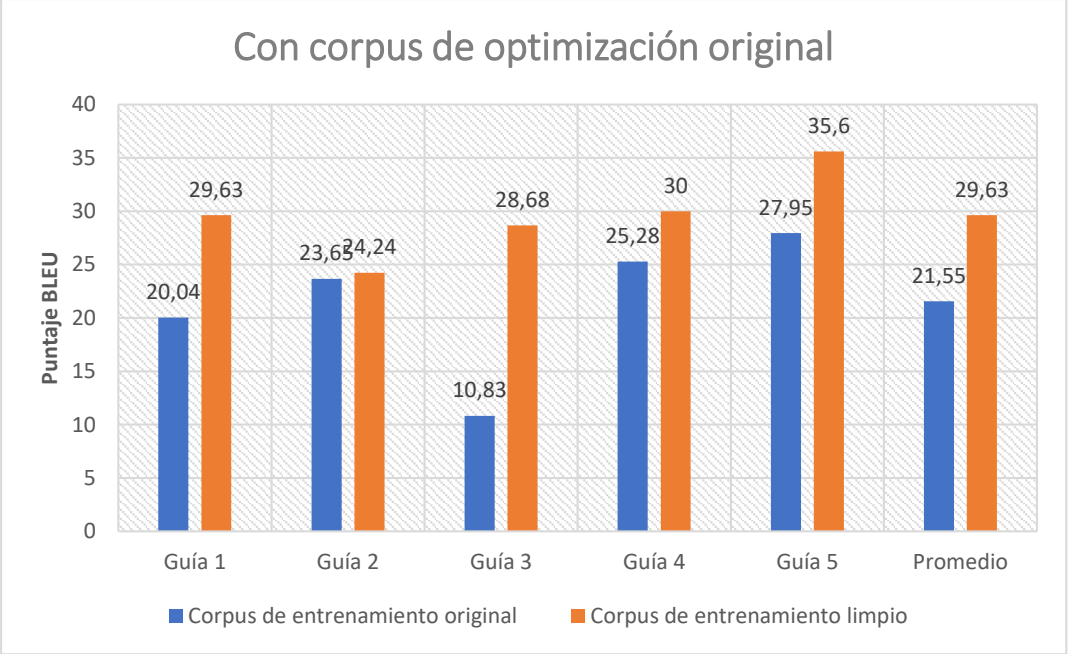


Figura 38. Puntajes BLEU de guías traducidas con el motor de MTradumática entrenado con el corpus de entrenamiento original vs. con el corpus de entrenamiento limpio, con corpus de optimización original.

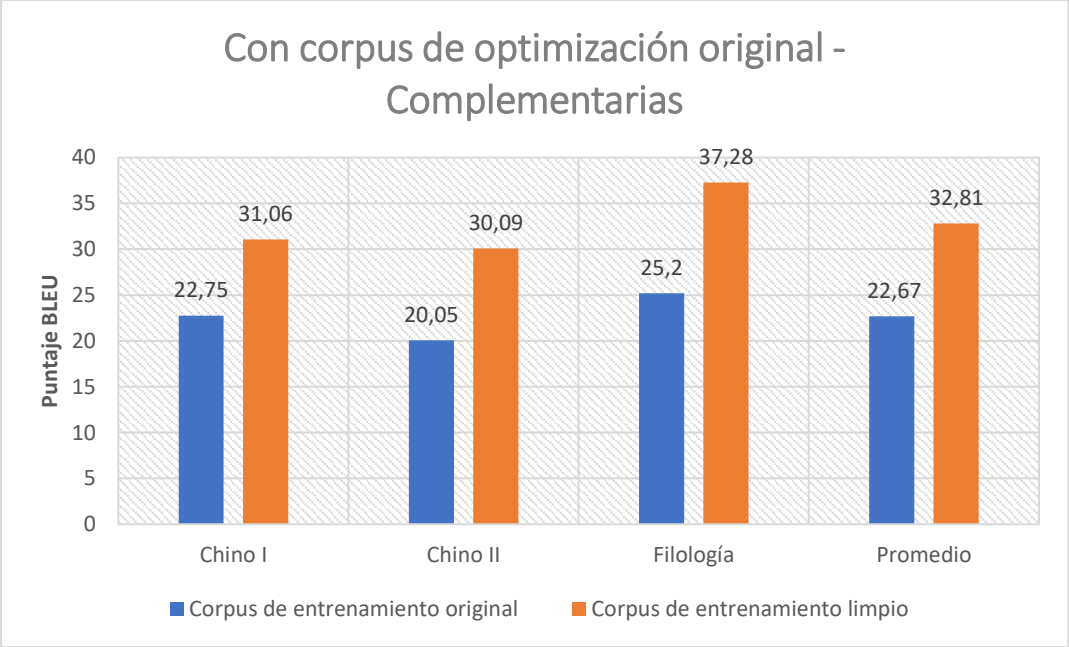


Figura 39. Puntajes BLEU de las guías complementarias traducidas con el motor de MTradumática entrenado con el corpus de entrenamiento original vs. con el corpus de entrenamiento limpio, con corpus de optimización original.

Tanto en la figura 38, correspondiente a las traducciones de las guías originales, como en la figura 39, que corresponde a las traducciones de las guías complementarias, se observa que los motores de MTradumática que utilizan el

corpus de optimización original arrojan resultados de calidad muy baja. El mayor puntaje no llega a 38, lo que indica que un corpus de optimización sin editar no sólo no mejora la calidad de las traducciones automáticas, sino que puede incluso perjudicarla. Si bien aquí gana el corpus de entrenamiento limpio, los puntajes son tan bajos en general que se puede afirmar que los motores ilustrados en las figuras 37 y 38 producen traducciones automáticas inviables y, por lo tanto, se pueden descartar.

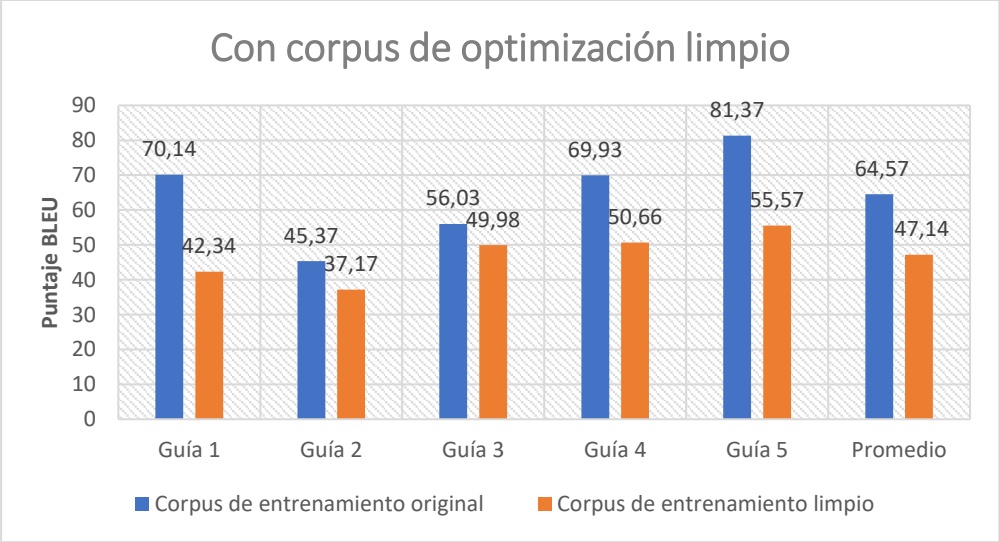


Figura 40. Puntajes BLEU de guías traducidas con el motor de MTradumática entrenado con el corpus de entrenamiento original vs. con el corpus de entrenamiento limpio, con corpus de optimización limpio.

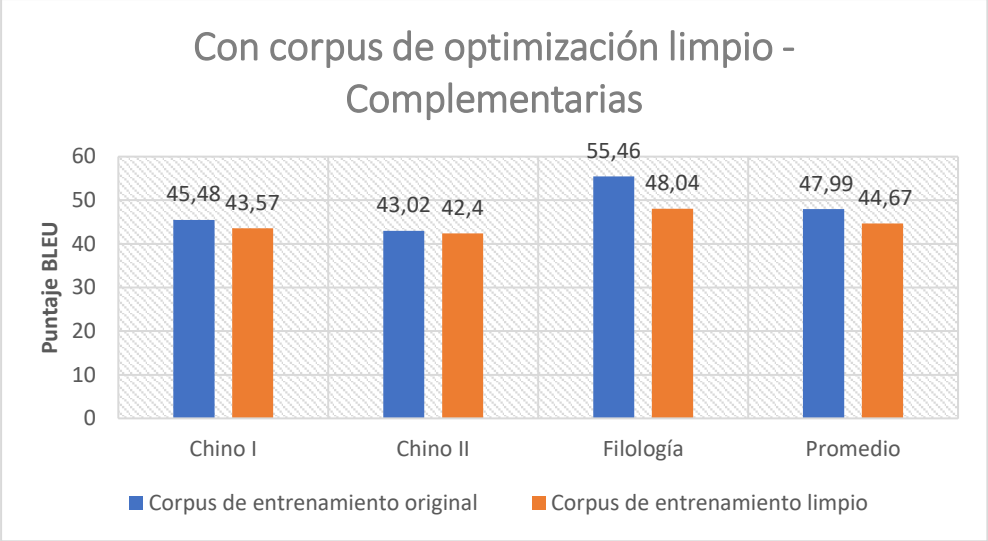


Figura 41. Puntajes BLEU de las guías complementarias traducidas con el motor de MTradumática entrenado con el corpus de entrenamiento original vs. con el corpus de entrenamiento limpio, con corpus de optimización limpio.

En estos motores se vuelve a ver la preponderancia del corpus de entrenamiento original sobre el corpus de entrenamiento limpio. En la figura 40, correspondiente a las guías originales, la diferencia entre los dos motores oscila entre 6,05 y 27,8 puntos a favor del primero. Por otro lado, la figura 41 muestra una diferencia más modesta, entre 0,62 y 7,42 puntos a favor del motor entrenado con el corpus de entrenamiento original; sin embargo, ya que la limpieza del corpus de entrenamiento implica un proceso extra, sigue siendo significativa.

Teniendo en cuenta tanto las traducciones de las guías originales como las de las guías complementarias, los datos apuntan a que, tratándose del corpus de entrenamiento, la cantidad es mucho más importante que la calidad.

6.4. Corpus de optimización: ninguno vs original vs limpio

En este apartado se interpretará la influencia del corpus de optimización en los puntajes BLEU obtenidos por los diferentes motores de MTradumática.

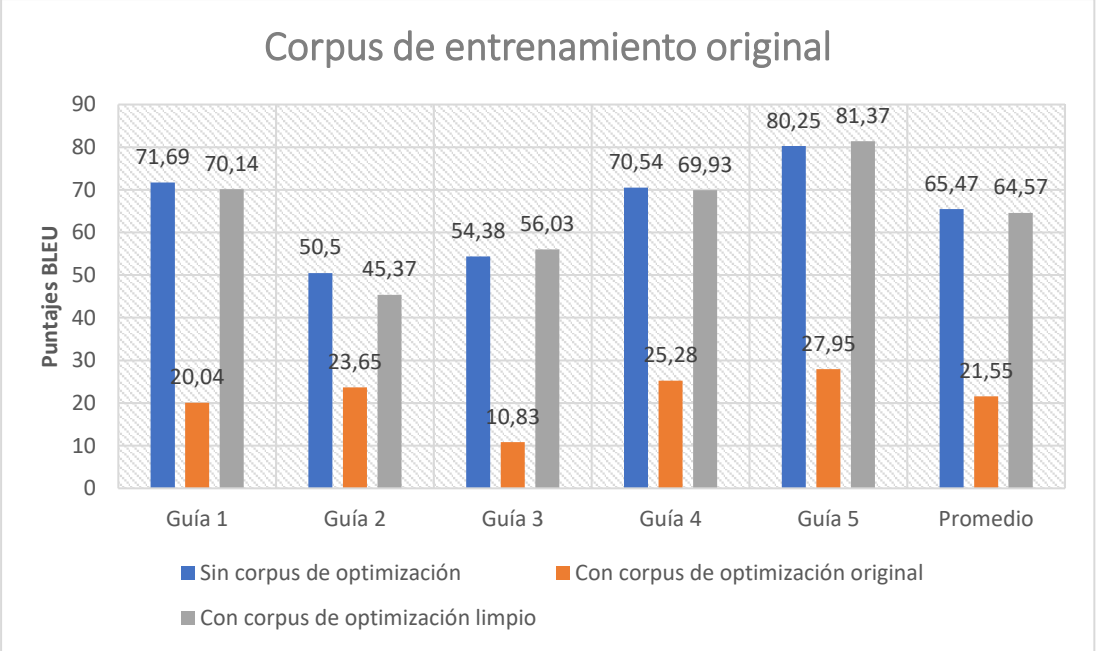


Figura 42. Puntajes BLEU de guías traducidas con el motor de MTradumática optimizado con el corpus de optimización original vs. con el corpus de optimización limpio, ambos entrenados con el corpus de entrenamiento original.

A simple vista se aprecia en la figura 42 que los motores entrenados con el corpus de optimización original producen traducciones de una calidad severamente inferior a la de los entrenados con el corpus de optimización limpio y los entrenados sin ningún corpus de optimización, a tal punto de que sus puntajes son dos o tres categorías mas bajos. Por otro lado, la diferencia entre los motores entrenados sin corpus de optimización y los entrenados con el corpus de optimización limpio, aunque existe, es muy leve, y puede favorecer a una u otra de las opciones.

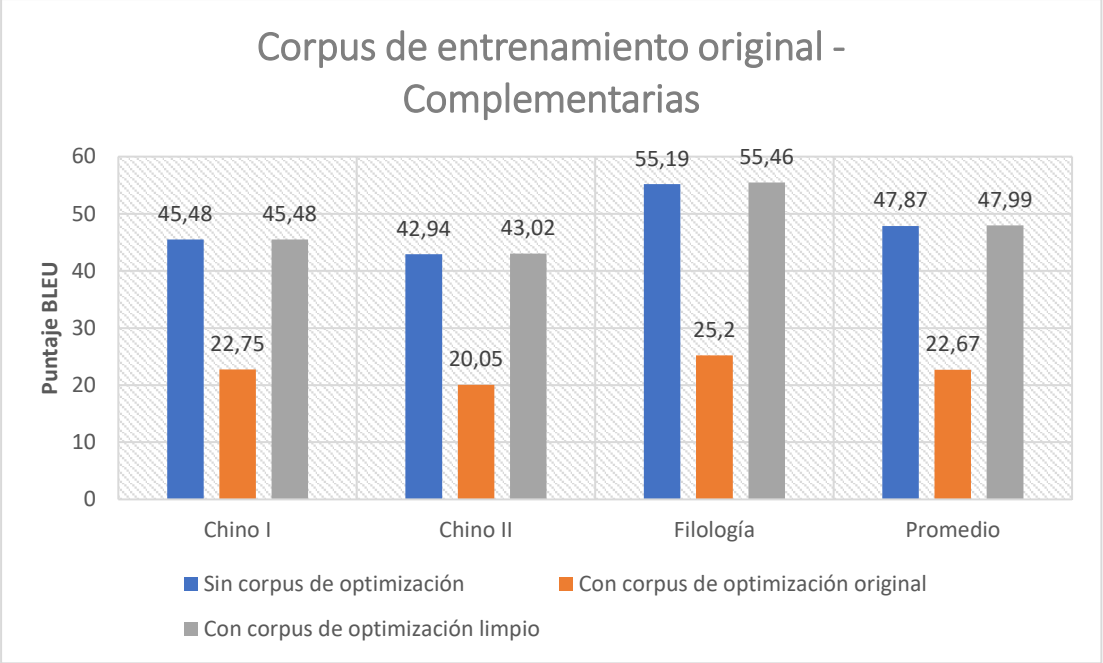


Figura 43. Puntajes BLEU de las guías complementarias traducidas con el motor de MTradumática optimizado con el corpus de optimización original vs. con el corpus de optimización limpio, ambos entrenados con el corpus de entrenamiento original.

Al observar los puntajes correspondientes a las guías complementarias, aunque la diferencia es algo menor, sigue siendo notoria. Por otro lado, los puntajes de los motores entrenados sin corpus de optimización y los entrenados con corpus de optimización limpio son iguales, o tienen una diferencia insignificante.

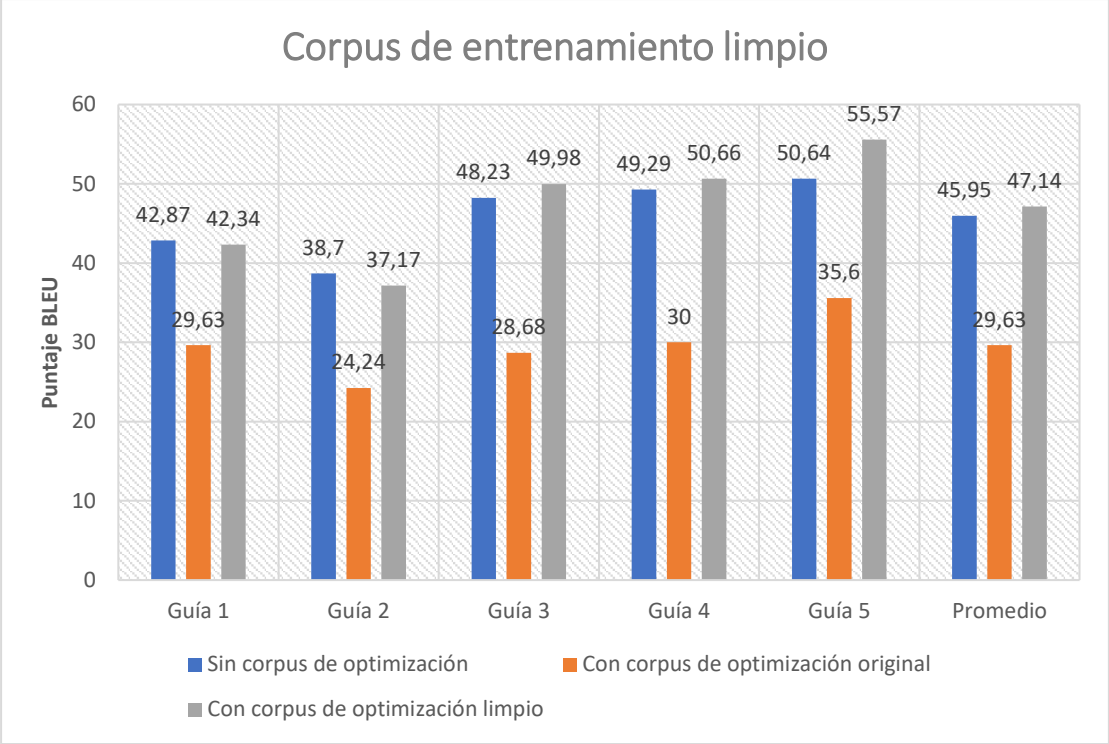


Figura 44. Puntajes BLEU de guías traducidas con el motor de MTradumática optimizado con el corpus de optimización original vs. con el corpus de optimización limpio, ambos entrenados con el corpus de entrenamiento limpio.

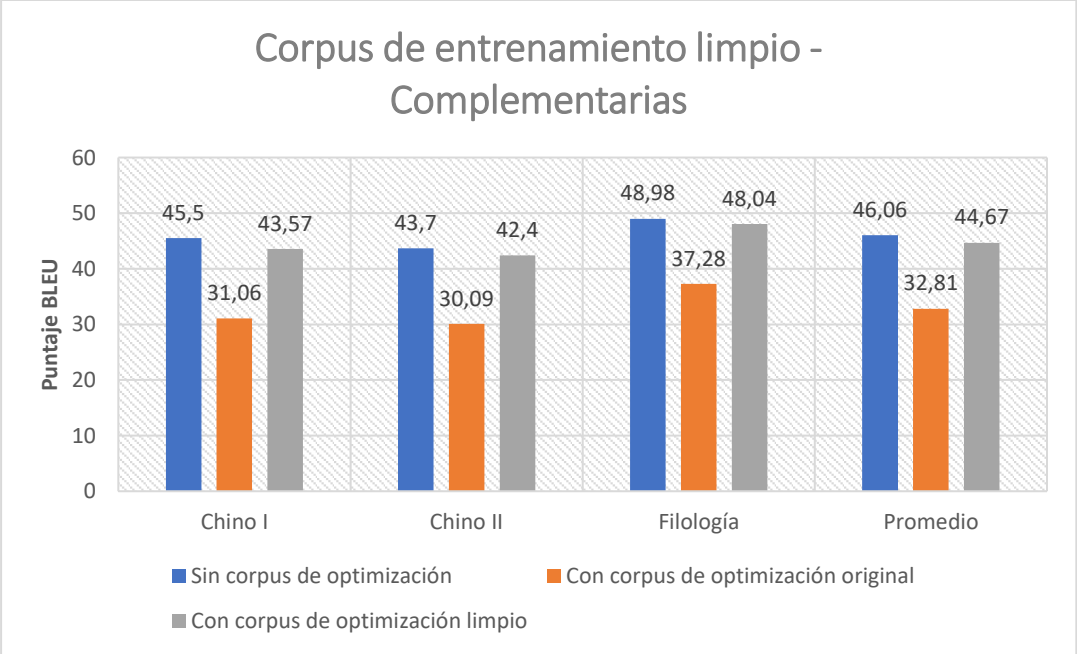


Figura 45. Puntajes BLEU de las guías complementarias traducidas con el motor de MTradumática optimizado con el corpus de optimización original vs. con el corpus de optimización limpio, ambos entrenados con el corpus de entrenamiento limpio.

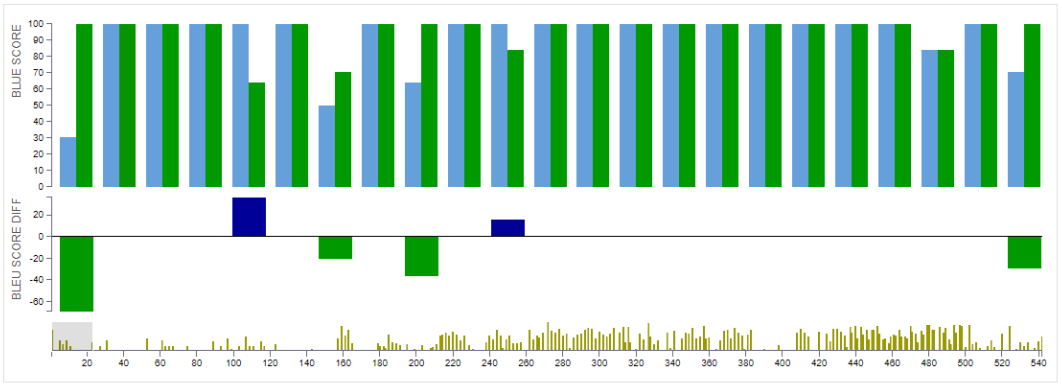
Las figuras 44 y 45 ilustran los puntajes de las traducciones de las guías originales y las guías complementarias, respectivamente, hechas con los motores de

MTradumática entrenados con el corpus de entrenamiento limpio. Al igual que en el caso de los motores entrenados con el corpus de entrenamiento original, aquí se aprecia la disminución en la calidad de las traducciones (tal como la evalúa el sistema BLEU) realizadas con los motores entrenados con el corpus de optimización original, comparadas con las demás. También se observa la cercanía entre los puntajes de los motores entrenados con corpus de optimización limpio y los entrenados sin corpus de optimización en absoluto.

A partir de estos resultados se puede deducir que, a diferencia del caso del corpus de entrenamiento, cuando se trata de la optimización de los motores de traducción automática, la calidad sí importa más que la cantidad. Es preferible no optimizar el motor de traducción automática que hacerlo con un corpus que contenga demasiados errores. En cuanto a si incluir un corpus de optimización limpio o no incluir ningún corpus de optimización en lo absoluto, los resultados no permiten realizar una afirmación contundente según el sistema de evaluación BLEU.

6.5. Comparación cualitativa

El sistema de evaluación BLEU es un método cuantitativo para calificar las traducciones producidas por los motores de traducción automática. Sin embargo, su alcance puede ser algo estrecho y obviar algunos factores importantes a la hora de decidir qué motor de traducción automática usar para traducir las guías docentes del área de humanidades de la UAB, o cómo mejorar los procesos de entrenamiento de los motores. En este apartado se señalan algunos ejemplos encontrados en la comparación de traducciones de la guía 5.



<< < 209 > >> From 1 to 23 of all **542** sentences.

Sentence 209	BLEU	Length ratio	Text
Source	-	-	Se excluyen de la recuperación las actividades de evaluación en que se hayan producido irregularidades.
Human	100.00	1.00	evaluation activities in which such irregularities have occurred are excluded from second - chance examination .
Machine	3.26	0.63	excluding recovery assessment activities that there have been irregularities .
Machine	15.79	0.75	students may not retake assessment activities in which irregularities have occurred .

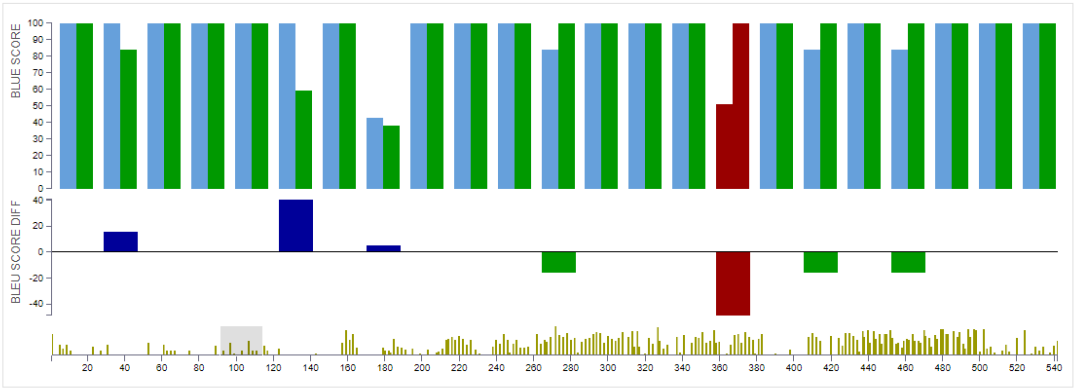
Figura 46. Evaluación BLEU de un segmento.

Traducción humana: *evaluation activities in which such irregularities have occurred are excluded from second - chance examination.*

Traducción con motor de MTradumática: *students may not retake assessment activities in which irregularities have occurred. (puntaje BLEU: 15.79)*

Los dos segmentos son bastante diferentes en cuanto a su longitud, la escogencia y orden de las palabras, pero ambos son coherentes, fluidos, y expresan la misma idea, aunque de diferente manera. Esto no se ve reflejado en el puntaje dado por el sistema BLEU.

Versus



<< < 107 > >> From 92 to 114 of all **542** sentences.

Sentence 107	BLEU	Length ratio	Text
Source	-	-	Preparación de ejercicios
Human	100.00	1.00	preparation of exercises
Machine	51.00	0.67	preparation exercises
Machine	100.00	1.00	preparation of exercises

Figura 47. Evaluación BLEU de un segmento adicional.

Traducción humana: *preparation of exercises*

Traducción de Google Translate: *preparation exercises* (puntaje BLEU: 51.00)

En este caso, la diferencia es de sólo una palabra (of). Sin embargo, esa palabra provoca un cambio semántico importante en los segmentos: la oración *preparación de ejercicios* cambia a *ejercicios de preparación*.

	A	B	C	D	E	F	G
20	19	Actividades	activities	activities	activities	100.000.000	100.000.000
21	20					100.000.000	100.000.000
22	21	Evaluación	assessment	evaluation	evaluation	84.089.642	84.089.642
23	22					100.000.000	100.000.000
24	23	Actividades de evaluación	assessment activities	evaluation activities	assessment activities	70.710.678	100.000.000
25	24					100.000.000	100.000.000
26	25	Bibliografía	bibliography	bibliography	bibliography	100.000.000	100.000.000
27	26					100.000.000	100.000.000
28	27	Titulación	degree	title	degree	84.089.642	100.000.000
29	28					100.000.000	100.000.000
30	29	Tipo	type	type	type	100.000.000	100.000.000
31	30					100.000.000	100.000.000

Figura 48. Comparación entre traducciones de la guía 5 realizadas con Google Translate y con motor entrenado con MTradumática

En los segmentos 23 y 27, ilustrados en la Figura 48, se observa que el sistema BLEU disminuye el puntaje de calificación por el uso de términos que no corresponden exactamente a los utilizados en la traducción humana, así sean sinónimos y no afecten la inteligibilidad de la traducción. Es decir, ya que se basa estrictamente en n-gramas, no reconoce el sentido completo de un segmento ni

necesariamente distingue sinónimos o entre diferencias meramente estilísticas y errores que afecten la inteligibilidad o el sentido de un texto.

7. CONCLUSIONES

En el siguiente apartado se explicarán las conclusiones a las que se ha llegado a través de este trabajo.

La primera conclusión es que la decisión de cuál herramienta es superior entre MTradumática y Google Translate a la hora de traducir documentos especializados como, por ejemplo, las guías docentes de la UAB, depende de quien realice las traducciones humanas.

La segunda conclusión importante es que en cuanto se refiere al corpus de entrenamiento, la cantidad es más importante que la calidad. Para que el motor produzca traducciones con una mejor calificación no es necesario ocupar tiempo en limpiar el corpus de entrenamiento, sino en agrandar el corpus existente, agregando constantemente la mayor cantidad de segmentos posible. En este sentido, la primera parte de la hipótesis de este trabajo queda negada.

El corpus de optimización, por otro lado, tal como se plantea en la hipótesis de este trabajo, sí depende de la calidad, a tal punto que es preferible omitir este paso en el entrenamiento del motor de traducción automática que utilizar un corpus inadecuado, ya que al hacerlo el motor puede pasar de crear traducciones de buena calidad a crear unas prácticamente inutilizables.

También se concluye que bien el sistema de evaluación BLEU sirve para indicar si una traducción automática tiene una calidad base suficiente o no, la puntuación no indica la intensidad de los errores. Es decir, no distingue entre diferencias estilísticas y errores gramaticales o semánticos que puedan comprometer la inteligibilidad del texto. Por esta razón es esencial la evaluación y/o la posedición humana para identificar aquellos matices. Al menos en el ámbito de la traducción especializada, el

ser humano sigue siendo el socio de la traducción automática, tal como lo señalaba Bar-Hillel en 1959.

8. RECOMENDACIONES

En este apartado 8.2. se harán recomendaciones a partir de las conclusiones derivadas del presente trabajo a la hora de usar motores de traducción automática de MTradumática:

1) Cuando se trata del corpus de entrenamiento, priorizar la cantidad sobre la calidad. Entre más segmentos tenga el corpus, mejor, independientemente de su nivel de corrección.

2) Limpiar y corregir cuidadosamente el corpus de optimización, de tal manera que contribuya al perfeccionamiento y no al empeoramiento de los motores TAE creados con MTradumática.

3) Redactar en las guías originales toda la información, incluyendo aquella que es exclusiva para los estudiantes hispanoparlantes, y la que es exclusiva para los estudiantes angloparlantes, usando en su lugar títulos y fuentes diferentes u otros mecanismos para diferenciar los contenidos.

4) Tener en cuenta el sistema de evaluación BLEU como un sistema de descarte más que como un sistema de confirmación de calidad. Es decir, que sirve para detectar los motores que no funcionan, pero es menos confiable a la hora de detectar los mejores motores.

5) Confirmar la corrección de las guías docentes en español, para aumentar la precisión de las traducciones automáticas, sea cual sea el motor TAE que se use.

6) Realizar más estudios para establecer los mejores motores de traducción automática y los mejores sistemas de evaluación de motores TAE para el caso de las guías docentes de la UAB.

9. BIBLIOGRAFÍA

BIBLIOGRAFÍA

Aranberri, N. (2014). Posedición, productividad y calidad. *Tradumàtica*, *12*, 0471-0477. https://doi.org/10.5565/rev/tradumatica.62

Babych B. (2014). Mètriques d'avaluació automatitzada de TA i les seves limitacions. *Revista Tradumàtica: tecnologies de la traducció*, *0*(12), 464-470. https://doi.org/10.5565/rev/tradumatica.70

Bar-Hillel, Y. (1951). The present state of research on mechanical translation. *American Documentation*, *2*(4), 229-237. https://doi.org/10.1002/asi.5090020408

Dogru, G., Martín-Mor, A., & Aguilar-Amat, A. (2018, mayo 10). *Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora*.

EAMT | European Association for Machine Translation. (s. f.). Recuperado 13 de agosto de 2020, de http://www.eamt.org/mt.php

Evalúa modelos | Documentación de AutoML Translation. (s. f.). Google Cloud. Recuperado 12 de agosto de 2020, de https://cloud.google.com/translate/automl/docs/evaluate?hl=es-419

Fernández Ruiz, M. E., & Sánchez-Gijón, P. (2019). Entrenamiento y comparativa de motores de TAE especializados en la localización de aplicaciones móviles. *Tradumàtica*, *17*, 0162-0183. https://doi.org/10.5565/rev/tradumatica.229

Forcada, M. L. (2009). Apertium: Traducció automàtica de codi obert per a les llengües romàniques. *Linguamática*, *1*, 13-23.

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

Green, S., Heer, J., & Manning, C. D. (2013). *The Efficacy of Human Post-Editing for Language Translation*. 10.

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

Hutchins, J. (s. f.). *Machine translation: A concise history*. 21.

ISO 18587:2017(en), Translation services—Post-editing of machine translation output—Requirements. (2017). https://www.iso.org/obp/ui/#iso:std:iso:18587:ed-1:v1:en

KantanMT. (2015, enero 21). *Moses Use Case: KantanMT.com*. https://kantanmtblog.com/2015/01/21/moses-use-case-kantanmt-com/

Kelly, N. (2011, noviembre 6). *Ray Kurzweil on Translation Technology*. https://vimeo.com/25021517

Koehn, P. (2020). *Statistical Machine Translation System User Manual and Code Guide*. http://www.statmt.org/moses/manual/manual.pdf

Martín Mor, A. (2017). MTradumàtica: Statistical machine translation customisation for translators. *Skase. Journal of Translation and Interpretation*, *11*(1), 0025-0040.

Martín-Mor, A. (2017). MTradumàtica: Statistical Machine Translation Customisation for Translators. *SKASE Journal of Translation and Interpretation*, *11*, 25-39.

Martin-Mor A., & Piqué i Huerta R. (2017). MTradumàtica i la formació de traductors en Traducció Automàtica Estadística. *Revista Tradumàtica: tecnologies de la traducció*, *0*(15), 97-115. https://doi.org/10.5565/rev/tradumatica.199

Melby, A. K. (2012). *HUMAN AND MACHINE TRANSLATION QUALITY: DEFINABLE? ACHIEVABLE? DESIRABLE?* 29.

Nagao, M. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. En A. Elithorn & R. Banerji (Eds.), *Artificial and Human Intelligence*. Elsevier Science Publishers. B.V.

Nunes Vieira, L. (2019). Post-Editing of Machine Translation. En M. O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology* (pp. 319-335).

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

Parra, C. (2018, abril). *Evolución de la traducción automática. La Linterna del Traductor, n.º 16*. http://www.lalinternadeltraductor.org/n16/traduccion-automatica.html

Pautas para la postedición de la traducción automática. (2013, septiembre 23). TAUS - The Language Data Network. https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines-spanish

Ping, K. (2009). Machine Translation. En M. Baker & G. Saldanha (Eds.), *Routledge encyclopedia of translation studies* (2.ª ed., Vol. 2, pp. 162-168). Routledge.

Pym, A. (2019). Quality. En M. O'Hagan (Ed.), *The Routledge Handbook of Translation Technology* (pp. 437-452). Routledge.

Sánchez-Gijón, P. (2016). La posedición: Hacia una definición competencial del perfil y una descripción multidimensional del fenómeno. *Sendebar*, *27*(0), 151-162-162. https://doi.org/10.30827/sendebar.v27i0.4016

Sánchez-Martínez, F. (2012). Motivos del creciente uso de traducción automática seguida de posedición. *Tradumàtica*, *10*, 0150-0156. https://doi.org/10.5565/rev/tradumatica.24

Translation services—Service requirements. (s. f.). Recuperado 14 de agosto de 2020, de https://images10.newegg.com/UploadFilesForNewegg/itemintelligence/ACCO/prEN_150381519719906828.pdf

Understanding MT Quality: BLEU Scores. (s. f.). SDL. Recuperado 12 de agosto de 2020, de https://blog.sdl.com/blog/understanding-mt-quality-bleu-scores.html

What is MTradumàtica? (s. f.). Mtradumàtica. Recuperado 13 de agosto de 2020, de https://tradumatica.github.io/

10. ANEXOS

Archivos guías originales: https://bit.ly/31187CL

Archivos guías complementarias: https://bit.ly/3lE0N7V

Corpus de entrenamiento y optimización: https://bit.ly/3118n4H

Tablas de puntajes BLEU: https://bit.ly/3doDZ9A