La traducción automática en la literatura: creación de un corpus paralelo inglés-español para entrenar y evaluar un sistema neuronal en el marco del proyecto MTUOC

Leonor Moreno Pons

Trabajo Final de Máster
Tutor/a: Dr./Dra. Silvia Rodríguez Vázquez
Máster de Traducción y Tecnologías
Universitat Oberta de Catalunya
Junio de 2022

Resumen

Los nuevos y significativos avances tecnológicos en traducción automática han contribuido a su papel decisivo en distintas modalidades de los servicios de traducción. En el contexto literario, sin embargo, el uso de la traducción automática suscita bastantes reticencias debido a las características propias de los textos literarios ubicados hasta hoy en un terreno bastante inexplorado para la innovación tecnológica en traducción. En los últimos años, con el fin de superar el difícil encaje entre traducción automática y literatura se ha prestado atención a la investigación sobre la creación de sistemas automáticos específicos de traducción literaria. Con la mirada puesta en investigaciones previas y con el apoyo necesario de las herramientas y programas de código libre del proyecto MTUOC, este trabajo se centra en la elaboración de un corpus paralelo de temática literaria destinado a entrenar un sistema de traducción automática neuronal para traducir textos de ficción. El proceso no ha estado exento de dificultades y obstáculos, los cuales nos han permitido identificar la problemática específica de un corpus creado a partir de obras literarias, en particular, la insuficiente precisión de las técnicas automáticas de alineación para garantizar un rendimiento óptimo del motor. El estudio termina con una evaluación automática comparativa entre nuestro motor entrenado y un sistema generalista y potente como Google Translate. A pesar de los problemas, las mediciones muestran resultados similares y perfectamente comparables entre ambos sistemas. A su vez, se describen posibles líneas de investigación futura para superar las dificultades y conseguir mejores resultados que favorezcan la aceptación de la traducción automática como una herramienta útil en la traducción literaria.

Palabras clave: corpus paralelo alineado, entrenamiento de sistemas de traducción automática, traducción literaria, evaluación automática

Abstract

Recent and relevant technological advances have consolidated machine translation as an important player in different types of translation services. In the literary context, however, the use of machine translation raises reluctance due to the special features of literary texts which are so far placed in a quite unexplored terrain for technological innovation in machine translation. In recent years, in order to get over the gap between machine translation and literature, attention has been paid to research on the creation of specific machine systems of literary translation. With an eye on previous research and the support of the open source tools and programs of the MTUOC project, this work focuses on the creation of a parallel corpora of literary content aimed at training a neural machine translation engine to translate fictional texts. The procedure has not been without difficulties and obstacles which have allowed us to identify the specific problems of a literary based-corpus, such as the insufficient precision of automatic

alignment techniques to guarantee optimal engine performance. The study ends with a comparative automatic metrics evaluation between our trained engine and a general and powerful system such as Google Translate. In spite of the problems, the metrics show similar and perfectly comparable results between both systems. In addition, possible lines of further research are described to overcome the difficulties and achieve better results that build confidence in machine translation as a useful tool in literary translation.

Keywords: aligned parallel corpora, machine translation engine training, literary translation, automatic metrics evaluation

Resum

Els darrers i significants avenços tecnològics en traducció automàtica han contribuït al seu paper decisiu en diverses modalitats dels serveis de traducció. En el context literari, però, l'ús de la traducció automàtica suscita força reticències a causa de les característiques pròpies dels textos literaris ubicats a hores d'ara en un terreny força inexplorat per a la innovació tecnològica en traducció. En els darrers anys, per tal de contribuir a superar el difícil encaix entre traducció automàtica i literatura s'ha prestat atenció a la investigació sobre la creació de motors automàtics específics de traducció literària. Amb la mirada posada en investigacions prèvies i amb el suport necessari de les eines i els programes de codi lliure del projecte MTUOC, aquest treball se centra en l'elaboració d'un corpus paral·lel de contingut literari destinat a entrenar un sistema de traducció automàtica neuronal per traduir textos de ficció. El procés no ha estat lliure de dificultats i obstacles, els quals ens han permès identificar la problemàtica específica d'un corpus basat en obres literàries, com ara la insuficient precisió de les tècniques automàtiques d'alineació per garantir un rendiment òptim del motor. L'estudi acaba amb una avaluació automàtica comparativa entre el nostre motor entrenat i un sistema generalista i potent com Google Translate. Malgrat les dificultats, les mètriques mostren resultats similars i perfectament comparables entre els dos sistemes. Alhora, es descriuen possibles línies d'investigació futura per superar les dificultats i aconseguir millors resultats que recolzen la traducció automàtica com una eina útil en traducció literària.

Paraules clau: corpus paral·lel alineat, entrenament de sistemes de traducció automàtica, traducció literària, avaluació automàtica

Tabla de contenido

1. Introducción

1.1 Contexto de trabajo

El presente trabajo final de máster (TFM) en Traducción y Tecnologías se enmarca en el campo de investigación de las tecnologías de la traducción; concretamente, en el ámbito de la creación de corpus paralelos, su aplicación a la traducción automática (TA) y, más específicamente, la traducción de textos de ficción literaria.

El desarrollo y la creciente utilización de la tecnología en la traducción está revolucionando las bases tradicionales de la experiencia traductora tanto en el ámbito formativo como profesional e investigador. La traducción automática es posiblemente el ámbito que más ha incrementado su presencia en la industria de la traducción como demuestra la gran cantidad de estudios dedicados a esta nueva realidad desde diversos puntos de vista; desde los más técnicos, dedicados a perfeccionar con avances continuos los sistemas de TA (Forcada, 2017; Vaswani & et al., 2017) hasta los dedicados a estudiar su impacto en la práctica traductora (Koponen et al., 2019; Oliver, A. & Alvarez-Vidal, 2021) o los que analizan y evalúan la calidad de los resultados (Görög, 2014; Koby et al., 2014; Toral & Way, 2018).

Si bien en ámbitos especializados el uso de la TA está cada vez más extendido (Pym, 2013; Serrano Vega, 2015), en el terreno de la literatura son muchos los recelos que despierta entre los traductores, como hemos podido comprobar en varios trabajos (Serrano, 2020; Guerberof & Toral, 2022) No nos ocuparemos aquí de la postedición como método que complementa la TA y favorece su percepción, ni tampoco trataremos de rebatir los argumentos de los profesionales que, en la mayoría de los casos son fruto de su experiencia y muchas veces del temor a verse desplazados.

En el presente trabajo, cuyo enfoque es fundamentalmente técnico, nos sumergimos en la aplicación de las tecnologías para crear instrumentos de trabajo como son un corpus paralelo y un motor de traducción automática. La otra vertiente de nuestro proyecto se desarrolla en el ámbito de la traducción literaria; es decir, tanto nuestro corpus como el motor entrenado a partir del mismo están destinados a traducir textos de contenido literario, un terreno en el que, como veremos, se están invirtiendo muchos recursos de

investigación por las posibilidades que ofrece combinar dos campos que, a priori, parecen incompatibles (Serrano, 2020).

1.2 Motivación

Uno de los aspectos que más nos han interesado en el transcurso de este máster ha sido el de la investigación en tecnologías aplicadas a la traducción. Ya sea por lo novedoso del tema o por la convicción de la utilidad de este tipo de herramientas, nuestro punto de partida ha sido el de mirar a la máquina como un colaborador y no como una imposición.

Como veremos con detalle al analizar el proceso de traducción de un texto literario (Landers, 2001; Song Xiaoshu, 2003; Udina, 2018; Guerberof &Toral, 2022), el traductor se preocupa por proporcionar una experiencia lectora similar a la del lector original; de esta manera, ante los problemas de traducción, se generan y barajan diversas soluciones posibles de entre las cuales el traductor elige la mejor opción según el texto, el contexto y cultura de destino. En esta estrategia, pensamos que los resultados proporcionados por la TA tomados como sugerencias —que pueden ser aceptadas, rechazadas o modificadas— pueden ser de gran ayuda a la hora de tomar decisiones.

En este sentido, compartimos la idea de que la TA —si proporciona resultados aceptables en términos de calidad (Briva-Iglesias, 2020)— garantiza la fidelidad de la traducción y puede ser un buen punto de partida para el traductor literario para lograr el objetivo de transmitir el mensaje y el estilo del original.

Por todo ello, la creación de un corpus paralelo literario y su aplicación al montaje de un motor de TA específicamente preparado para traducir literatura nos pareció una forma de tender un puente entre estas dos modalidades de traducción además de ser un terreno en el que los investigadores demandan más estudios sobre su viabilidad (Vieira, 2019; Youdale, 2020).

1.3 Objetivos y metodología

Los objetivos que guiarán las fases del presente trabajo se pueden concretar de la siguiente manera:

1. Crear un corpus paralelo alineado de obras literarias para el par de lenguas inglés (EN) - español (ES).

2. Entrenar y poner en marcha un motor de traducción automática neuronal con la metodología del proyecto MTUOC, utilizando el corpus literario creado.

3. Evaluar mediante métricas automáticas los resultados de traducción del motor entrenado y compararlos con los resultados de un sistema genérico; concretamente, *Google Translate*.

Los dos primeros objetivos, de carácter técnico, expresan el propósito de llevar a cabo, con muestras reales, el proceso de diseño y producción de un corpus paralelo literario y de un motor de TA.

Las preguntas que nos planteamos son: ¿Qué dificultades se pueden encontrar en la elaboración de esta tarea? ¿Las técnicas que usamos de segmentación y alineación serán suficientes para garantizar una calidad aceptable que permita el entrenamiento y puesta en marcha del motor?

Con el objetivo 3 se pretende validar el trabajo técnico realizado; es decir, proporcionar evidencias de que nuestro motor de TA es capaz de ofrecer traducciones de calidad y compararlos con sistemas genéricos. En este sentido, podemos esperar que nuestro motor preparado específicamente para traducir textos literarios será capaz de proporcionar mejores resultados que los de *Google Translate.*

Como viene señalado en el título de este trabajo y veremos en el capítulo 3, la **metodología** empleada para la consecución de esos objetivos se enmarca en el proyecto de investigación y transferencia de tecnología MTUOC[1] de la Universitat Oberta de Catalunya (UOC) que gestiona y desarrolla Antoni Oliver y que tiene como objetivos

[1] https://xwiki.recursos.uoc.edu/wiki/mat00001ca/view/MicroMooc%20Integraci%C3%B3n%20de%20traducci%C3%B3n%20autom%C3%A0tica%20neuronal%20en%20proyectos%20de%20traducci%C3%B3n%20con%20MTUOC/Tema%201.%20%C2%BFQu%C3%A9%20es%20MTUOC%3F/ Última consulta: 10 de abril de 2022

facilitar el entrenamiento y la integración, en entornos profesionales de traducción, de sistemas de TA estadística y neuronal garantizando a sus usuarios la confidencialidad de sus traducciones. El proyecto MTUOC distribuye bajo licencia libre GNU las herramientas para la creación, instalación y puesta en marcha de estos motores.

1.4 Estructura del trabajo

Este trabajo se inicia con un capítulo de revisión (cap.2) de trabajos destacados que han contribuido al desarrollo del campo de investigación y enmarcan nuestro proyecto. Son varios los ámbitos que atañen a este trabajo y, por tanto, hemos creído conveniente dividirlo en varios apartados o secciones. El apartado 2.1 está dedicado a revisar algunos trabajos que se ocupan del diseño y elaboración de corpus paralelos con aplicaciones en distintos terrenos como la investigación, la formación o la práctica traductora (González Rey, 2014; Molés-Cases, 2016; Nádvorníková, 2017; Leiva Rojo, 2018). La finalidad de nuestro corpus es el entrenamiento y puesta en marcha de un motor de traducción automática; por tanto, en el apartado 2.2 damos un breve repaso a la evolución de los sistemas de traducción automática y su funcionamiento.

Emplear la TA para traducir textos literarios es un tema muy discutido en el que entran en juego varios aspectos. En el apartado 2.3 analizaremos algunas características de la traducción literaria; en particular, el estilo y la creatividad como rasgo diferenciador de este tipo de traducción (Landers, 2001) y en el 2.3.1 nos detendremos en trabajos cuyo objeto de estudio es la traducción automática de textos literarios y su impacto en la práctica traductora (Toral & Way, 2015, 2018; Matusov, 2019). El capítulo termina con un apartado (2.4) dedicado a la evaluación de los resultados de la TA; analizaremos aquí los distintos criterios y métodos de evaluación que se pueden aplicar dependiendo de los aspectos que se quieran evaluar.

En el capítulo 3 se realiza un informe detallado de la metodología que hemos seguido en este trabajo. La primera fase (apartado 3.2) comienza con la descripción del diseño, las características del corpus y los materiales que han servido para elaborar el corpus literario. El propósito de entrenar un motor de TA exige que el tamaño del corpus final sea muy superior al del corpus literario (A. Oliver, 2020); por tanto, este cuerpo central literario ha sido combinado con un corpus de contenido general que hemos preparado

a partir de corpus paralelos de acceso libre con contenidos y temáticas afines; es decir, temas educativos, culturales, noticias, etc.

En el apartado 3.3 explicamos los pasos seguidos para el procesamiento, entrenamiento, montaje y puesta en marcha de un motor de TA con un sistema neuronal —Marian— alimentado con los datos del corpus preparado para traducir textos de ficción literaria.

El capítulo 4 está dedicado a informar sobre los retos y los resultados obtenidos de la investigación mediante una evaluación con métricas automáticas que, además de evaluar nuestros resultados, permite comparar nuestras traducciones con los resultados de un sistema generalista.

2. Revisión bibliográfica

2.1 El corpus lingüístico: definiciones y contexto

La lingüística del corpus es una rama de la lingüística que estudia la confección y análisis de **corpus lingüísticos**, los cuales podemos definir como una recopilación de muestras reales de una lengua que se reúnen bajo unos criterios específicos (Bowker & Pearson, 2002). En palabras de Zanettin "in the field of corpus linguistics, a corpus is by default assumed to be a collection of texts in electronic format which are processed and analyzed using software specifically created for linguistic research (Zanettin 2012, p.7)". Más que una disciplina, se trata de un enfoque metodológico que supone la utilización de corpus lingüísticos para realizar investigaciones basadas en datos obtenidos a partir del corpus. Son varias las disciplinas que han adoptado esta metodología de investigación y que utilizan algún tipo de corpus dependiendo de sus necesidades y objetivos. En el ámbito de la traducción, nos interesan sobre todo los corpus bilingües o multilingües; en concreto, los **corpus paralelos** que contienen textos y su traducción a una o varias lenguas. Se pueden distinguir varias modalidades de corpus paralelos, entre las que encontramos el **corpus alineado**, que es el resultado de un proceso de alineación al que se somete el corpus paralelo, en el cual se trata de encontrar correspondencias entre segmentos textuales que son traducciones equivalentes (Leiva Rojo, 2018). Los textos paralelos alineados se llaman también **bitextos,** un término propuesto por Harris (1988) que ya vio el potencial de estos corpus para crear automáticamente diversos recursos de traducción como diccionarios y bases terminológicas bilingües, ejemplos para el aprendizaje de idiomas por ordenador, las investigaciones de la lingüística contrastiva o las memorias de traducción para la práctica traductora.

2.1.1 La creación de corpus paralelos

Los corpus paralelos alineados representan y hacen explícita la conexión entre segmentos de lenguas distintas y, como hemos mencionado, esto supone un gran potencial en la creación de recursos de traducción. En este sentido, la digitalización de la información y la posibilidad de disponer de textos en formato electrónico que pueden ser procesados automáticamente ha incrementado enormemente la creación de estos

recursos. Zanettin (2014) explica que la creación de corpus paralelos robustos y fiables es una tarea exigente y laboriosa que comprende la búsqueda y compilación de textos traducidos, su conversión a un formato estándar (TXT) que permita su procesamiento y su alineación final.

Siguiendo el razonamiento de Leiva Rojo (2018), para elaborar un corpus debemos partir de la siguiente premisa: la finalidad del corpus determinará las características que deberá tener el corpus de manera que sea útil para el propósito para el que se constituye. Por su parte, González Rey (2014) en la descripción de la elaboración de un corpus literario paralelo diseñado como herramienta didáctica, enumera los requisitos que definen todo corpus: la *representatividad*, es decir, que la selección de textos sirva para caracterizar un estado de lengua; la *explicitación de los criterios* de composición, que ayuda a justificar la creación del corpus en función de la finalidad del proyecto; y la *coherencia* de los criterios con el proceso de recopilación y diseño.

A continuación, veremos cómo algunos autores explicitan los factores que definen los corpus paralelos en función de su finalidad, teniendo siempre en cuenta los requisitos que deben cumplir:

Costa Pellicer (2021) en su trabajo sobre la elaboración de un corpus especializado sobre la intolerancia a la lactosa señala la *cantidad* y la *calidad* como las dos características clave que permiten establecer su representatividad "con el fin último de facilitar la toma de decisiones ante los problemas de traducción que presentan esta tipología de textos" (ídem, p.3). En el caso de un corpus muy especializado como el que elabora la autora, primará la calidad sobre la cantidad en tanto que el corpus servirá para extraer información precisa y contrastada.

Por otro lado, Nádvorníková (2017) analiza los aspectos técnicos que influyen en el tratamiento y explotación del corpus paralelo *InterCorp,* un corpus de más de mil millones de palabras en 39 idiomas creado por el Instituto Nacional del Corpus Checo y compuesto en su parte central de temática literaria y científica. La autora cita el *tamaño* y la *composición* del corpus como los rasgos determinantes de la representatividad de los corpus lingüísticos y, en particular, concluye que el factor más importante que define los corpus paralelos es el de la *calidad de la alineación* que comienza por una buena segmentación de los textos, generalmente a nivel de frase. Las herramientas

informáticas de segmentación no siempre definen las fronteras de las oraciones de la misma manera para todas las lenguas y esto puede suponer un problema en la posterior alineación. El software de alineación para un corpus de las características de *InterCorp* tiene que ser automático, puesto que, debido a su gran tamaño, es inviable una alineación manual. *Hunalign* (Varga et al., 2005) es el programa de alineación automática utilizada para la construcción de este corpus, un programa "qui donne des résultats assez fiables" según Nádvorníková (2017, p.8) y que pueden ser mejorados con la intervención semi-manual de un editor de textos.

Zanettin (2014) argumenta que los corpus lingüísticos han tenido una influencia importante tanto en el campo de la investigación teórica como en la práctica de la traducción y han sido claves para el desarrollo de la tecnología necesaria para la traducción automática. En este terreno, un sistema de TA basado en corpus "does not presuppose linguistic knowledge but rather relies on very large-scale corpora as a source of data to produce new translations. […] These systems increase their accuracy as more parallel data allow for the implementation of better statistics" (ídem, p.15). Con esta premisa de la cantidad, el requisito de la alineación se lleva a cabo mediante técnicas automáticas siguiendo un modelo composicional en el que cada unidad de traducción contiene un segmento bi-textual lo más pequeño posible y sin ningún segmento unilateral. Estas técnicas, concluye Zanettin, pueden no ser suficientemente precisas para la alta calidad que requiere la investigación teórica, pero "automatic alignment techniques provide viable results for MT purposes" (ídem, p.18) en cuyo ámbito el tamaño sí importa.

2.1.2 El corpus literario

Toral & Way (2018) citan la aparición del libro electrónico —y la consecuente posibilidad de compilar obras literarias para ser procesadas por ordenador— como el acontecimiento tecnológico que ha posibilitado la creación de corpus paralelos especializados en literatura y su proyección en diversos campos de investigación en traducción.

Las dificultades de alineación que hemos visto para los corpus paralelos en general se acentúan en el caso de los corpus literarios. Para Simard (2020), la conexión entre las

dos partes de un segmento alineado es muchas veces una aproximación, ya que la motivación del traductor para elegir una palabra o expresión sobre otra posible se encuentra en un contexto más amplio que el de la simple correspondencia léxica. De la misma manera, los aspectos estilísticos y retóricos propios de los textos literarios son factores que estimulan la creatividad del traductor y, en muchos casos, el texto resultante puede llegar a ser irreconocible respecto al original (Santoyo, 1996). De esta circunstancia podemos deducir que la segmentación de los textos original y traducido no tendrá una correspondencia total; con lo cual, muchos segmentos pueden ser inservibles para la alineación.

A continuación, veremos distintos criterios que se han seguido para la creación de corpus literarios destinados a probar su rendimiento en traducción automática.

Toral & Way (2018) construyen un corpus paralelo relativamente grande con unas 133 novelas traducidas inglés/catalán que, antes de su procesamiento, constituían alrededor de 1 millón de pares de frases. Este corpus se combina con otros corpus generales paralelos como *Opensubtitles*. Además, usan corpus monolingües en la lengua de llegada (alrededor de 16 millones de frases) recogidos y compilados desde la red, cuya temática puede ser muy variada y no necesariamente de contenido literario. Con todo ello, se entrenan distintos sistemas de traducción automática que son evaluados y comparados con otros sistemas como veremos más adelante.

Por su parte, Matusov (2019) utiliza el sistema general de TA neuronal AppTek[2] ruso/inglés adaptándolo al contenido literario con una pequeña cantidad de pares de oraciones provenientes de libros de ficción: alrededor de 3000 segmentos alineados que forma el corpus literario especifico.

Nuestro propósito de entrenar un motor de TA para traducciones literarias ha determinado las características de nuestro corpus: hemos tratado de reunir la mayor cantidad posible de materiales traducidos de contenido literario, los cuales han constituido la parte central del corpus. Este núcleo central ha sido combinado con otros corpus paralelos alineados de temática variada, pero relacionada de alguna manera con la ficción literaria para poder entrenar un sistema de TA neuronal, que exige un gran

[2] https://www.apptek.com/ Última consulta 11 de mayo de 2022

volumen de datos. En la sección 3.2.1 daremos detalles de la composición final del mismo.

2.2 La traducción automática

Una de las aplicaciones más populares a día de hoy de los corpus paralelos alineados es la de entrenar un motor de traducción automática. Si repasamos brevemente el recorrido de la **traducción automática,** nos hemos de remontar al año 1950 cuando se pensó en aplicar y desarrollar software para traducir textos de una lengua a otra, pero tras los primeros logros, se abandonó esta línea de investigación por considerarla escasamente fiable y productiva. Más tarde, en los años 80, como explica Simard (2020) y coincidiendo con la aparición de los corpus paralelos en formato electrónico, los nuevos sistemas de TA experimentaron un gran auge y siguen en constante evolución y perfeccionamiento.

Seguiremos las explicaciones de Antoni Oliver[3] y Antonio Toral[4] para narrar brevemente en qué consisten los tres paradigmas principales de TA, y comprender mejor su evolución y el papel que juegan los corpus paralelos en la construcción de estos sistemas.

Las primeras estrategias utilizadas en traducción automática iban desde la traducción directa palabra por palabra a los **sistemas basado en reglas** que realizan un análisis de la lengua de partida y la de llegada antes de realizar la transferencia de una a otra; estos sistemas poseen como componentes principales los diccionarios (bilingües y monolingües), las reglas gramaticales y los analizadores morfológicos. Para su desarrollo es necesario el trabajo de un profesional experto que sepa combinar toda la información que proporcionan los componentes. En la actualidad, un ejemplo representativo de este sistema basado en reglas es Apertium[5], una plataforma de software libre que puede ser usada para construir motores de traducción automática para distintos pares de lenguas (Forcada et al., 2011).

[3] https://youtu.be/oa8xV6FxnfY Última consulta: 21 de mayo de 2022
[4] https://xwiki.recursos.uoc.edu/wiki/mat00001ca/view/Research%20on%20Translation%20Technologies/%20Machine%20translation/ Última consulta: 12 de mayo de 2022
[5] https://www.apertium.org/index.spa.html?dir=arg-cat#?dir=spa-eng&q= Última consulta: 12 de mayo de 2022

Con los **sistemas estadísticos** hubo un cambio de paradigma muy importante. Ya no se buscaba desarrollar sistema sofisticados para un par de lenguas con los recursos lingüísticos anteriores, sino que se trataba de entrenar un sistema válido para muchos pares de lenguas mediante un conocimiento lingüístico que ya se poseía; es decir, a través de corpus paralelos con pares de frases equivalentes formados a partir de muestras reales. El sistema utiliza técnicas estadísticas basadas en el cálculo de probabilidades sobre combinaciones de palabras *n-grams* para entrenarse; es decir, para aprender a traducir y decidir qué tipo de transformaciones necesita para una traducción correcta. Cuánta más información —corpus paralelos— posea el sistema, mejores resultados podrá proporcionar. Los distintos componentes de un sistema estadístico funcionan de la siguiente manera: cuando queremos traducir una frase, el sistema busca información en el corpus paralelo que lo alimenta y a partir de este *modelo de traducción* proporciona varios candidatos probables, a los cuales se les da una puntuación sobre cuál es la combinación y el orden más frecuente en un idioma determinado. Por último, el *modelo de lengua* se encarga de que la traducción sea fluida partiendo de la frecuencia de utilización en el idioma de destino y puntúa los distintos *outputs* proporcionados para que, finalmente, el sistema elija entre los distintos candidatos sopesando todas las puntuaciones para elegir la opción más correcta o más fluida.

Los motores de **traducción automática neuronal** basados en redes neuronales se introdujeron ya de manera teórica en los años noventa (Forcada & Ñeco, 1997), pero el hardware necesario para su implantación no se desarrolló hasta bien entrado el siglo XXI. Estos sistemas se siguen alimentando principalmente de corpus paralelos —también comparables—, pero su tratamiento es completamente distinto a los estadísticos. Su funcionamiento, tal y como explica Forcada (2017) se parece al de las neuronas de nuestro cerebro; de ahí su nombre. Las redes neuronales están compuestas por miles y miles de unidades cuya activación depende de los estímulos que reciba de otras unidades (neuronas) y del tipo de conexión que exista entre ellas. Si los sistemas estadísticos —y también los basados en reglas— representan las palabras como cadenas (g-a-t-o), en el sistema neuronal el entrenamiento consiste en la construcción de una única y gran red que utiliza una representación vectorial de las palabras. Las palabras se tratan como conceptos; es decir, como vectores numéricos con cientos de dimensiones

que el propio sistema, mientras aprende a traducir, también aprende a representar cada vez con mayor perfección. El sistema no sabe el significado ni el uso gramatical de las palabras, pero sí sabe qué palabras se usan en contextos similares y las agrupa en espacios cercanos; por ejemplo, los días de la semana por un lado o los colores por otro; los verbos de movimiento compartirían también una zona común y los números estarían representados en una zona del espacio parecida. Puesto que las palabras —los conceptos— tienen cientos de dimensiones en un sistema real, sabemos que no podemos visibilizar esa representación espacial, pero sí podemos entender que, con un hardware robusto, concretamente mediante potentes procesadores gráficos (GPU), un sistema neuronal será capaz de representar todas estas dimensiones y establecer relaciones entre ellas. Gracias a esta capacidad computacional, el sistema puede realizar operaciones matemáticas con los vectores numéricos que representan las palabras; por ejemplo, calcular la distancia entre dos palabras, y esto implica que será capaz de relacionar esta información y usarla para mejorar la traducción.

El método básico de sistema neuronal es el de la *arquitectura recurrente* que establece las conexiones —decodificar-codificar— palabra por palabra siempre mirando a la representación completa de la frase a traducir. A partir de aquí, han ido apareciendo varios métodos avanzados:

1) la *atención en la arquitectura recurrente*: implementa una extensión que decide a qué partes de la frase original se le presta más atención. Este modelo ya consiguió una mejora sustancial respecto a los sistemas estadísticos.

2) la *arquitectura transformer*, introducida por Vaswani et al. (2017) en la que no hay recurrencia palabra por palabra, sino que el modelo aprende a conectar las palabras por caminos más cortos. Este método ha conseguido resultados muy satisfactorios cuando la cantidad de datos paralelos que procesa es abundante. En este trabajo hemos entrenado un sistema *transformer* a partir del software de entrenamiento Marian[6] adaptado por el proyecto MTUOC.

3) por último, conviene citar los *modelos no supervisados* que solamente utilizan datos monolingües para aprender los patrones de traducción y están aportando resultados interesantes.

[6] http://www.aclweb.org/anthology/P18-4020 Última consulta: 14 de mayo de 2022

Las ventajas en el funcionamiento de los modelos neuronales sobre los estadísticos son (i) su poder de generalización al agrupar palabras con un mismo comportamiento en un mismo espacio multidimensional, (ii) su capacidad de generar modelos de lengua entre palabras distantes frente al modelo estadístico que funciona en cadenas de palabra por palabra y uniendo trozos de estas cadenas, (iii) las traducciones de sistemas neuronales son más fluidas e impredecibles frente a las estadísticas, que son más predecibles y mecánicas.

En cuanto a las desventajas, señalaremos que los sistemas neuronales necesitan un hardware potente y caro con procesadores gráficos GPU, mientras que un sistema estadístico se puede entrenar con un ordenador personal con CPU. Otra desventaja, que con el tiempo y los nuevos métodos se va subsanando, son los problemas de alucinación —traducciones sin sentido— y problemas de adecuación al significado, es decir, las traducciones pueden ser fluidas, pero no adecuarse al significado del original.

Por último, es necesario destacar el papel esencial del software libre en el desarrollo de estos sistemas de traducción automática. La mayoría de *Toolkits* —Moses, OpenNMT, Marian, etc.— poseen una licencia permisiva que no solo permite el acceso libre a los códigos y programas, sino que también favorece su desarrollo con contribuciones por parte de empresas, proyectos, investigadores, etc. Esto implica que la mayoría de sistemas funcionan con el sistema operativo Linux de código abierto.

Del mismo modo, el acceso a recursos lingüísticos libres ha contribuido enormemente al desarrollo de los sistemas de TA. Por ejemplo, corpus paralelos, diccionarios, enciclopedias, etc. que se pueden descargar gratuitamente de la red.

2.3 La traducción literaria

En un detallado estudio acerca de la traducción literaria, Landers (2001) analiza los esfuerzos y la gran dedicación que requiere traducir un texto literario. Por ejemplo, la decisión de si su trabajo debe o no parecer una traducción implicará por parte del traductor toda una serie de actuaciones a la hora de realizar su tarea: si debe primar la fluidez de la expresión o la fidelidad al original, si debe reflejar el estilo propio del autor o es factible que el traductor tenga su propio estilo, etc.

En esta línea, Song Xiaoshu (2003) discute el tema de la traducibilidad del estilo literario: "The translator should carefully appreciate the tone and spirit of the whole original work through words, sentences and paragraphs it is made up of" (ídem, p.2) y afirma que el principio que debe guiar la traducción literaria es *la reproducción del estilo del original*. Para lograrlo, el texto traducido debe parecerse de manera fiel, flexible y satisfactoria al texto original, tanto en el fondo como en la forma.

Por otro lado, los textos literarios incluyen aspectos formales y semánticos como las figuras retóricas, los juegos de palabras, el humor, la ironía, el contexto, etc. que conllevan un trabajo extra de interpretación y estimulan la creatividad del traductor literario (Serrano, 2020). Por tanto, el traductor literario se ve en necesidad de reflexionar y posicionarse sobre sus propias decisiones. En este sentido, Dolors Udina (2018) explica así algunos aspectos de su trabajo en su traducción de *Mrs. Dalloway* de V. Woolf al catalán:

> "Como para traducirlo tenía que entenderlo todo **en una de las fases de la traducción (pongamos en la tercera o cuarta revisión) tenía tendencia a explicitarlo todo, como mínimo para entenderlo yo.** En un momento dado me di cuenta de que, si lo que pretendía era hacerlo todo tan comprensible, perdía sin remedio la fuerza del texto al borrar los enigmas que la escritora dejó abiertos. Así, en una de las últimas lecturas (la octava o la novena) me dediqué a eliminar todo lo que había añadido para que se entendiera más. La fidelidad al texto original exigía este desnudamiento del sentido" (Udina, 2018, p.1)

Otro aspecto importante en las traducciones es el de la puntuación:

> "**Cuando empecé a traducir *La senyora Dalloway*, me confundía tanto punto y coma y buscaba una razón para eliminar al menos algunos.** No encontré la razón y, para seguir la música del original y ser fiel al estilo de la autora, me pareció que tenía que reproducir casi con exactitud su puntuación. Cuando consulto versiones francesas, sobre todo, me sorprenden los cambios que los traductores galos se permiten, incluso en el punto y aparte, por ejemplo, que creo que en general forman parte del estilo del autor o, al menos, de la manera como quiere contar la historia" (Udina, 2018, p.1)

Otro ejemplo significativo de la importancia de la puntuación es el de Lewis Carroll, que "utilizaba de forma extraña los **paréntesis**, las **mayúsculas** y la **letra cursiva**, […] una serie de **técnicas literarias** que buscaban captar la atención del lector para poder entender las cosas de formas alternativa" (Briceño V., 2021).

De estas reflexiones podemos extraer una idea que nos parecen relevante para apoyar el uso de la TA en traducción literaria: su capacidad para transmitir la historia tal y como está contada y evitar digresiones que cambian por completo el sentido original.

2.3.1 La traducción automática de textos literarios

Junto a la aparición del libro electrónico que posibilita la creación de corpus paralelos especializados en literatura, Toral & Way (2018) añaden los enormes avances en la traducción automática neuronal —con un nivel de calidad cada vez mayor— como el contexto necesario para que los investigadores entraran en un terreno hasta hace poco inexplorado: el de la aplicación de técnicas computacionales a la traducción literaria.

Si bien la TA se ha aplicado con bastante éxito y aceptación a la traducción especializada porque asegura la fidelidad al original y facilita la coherencia terminológica, es más difícil encontrar ejemplos y apoyos de su utilización para la traducción literaria. Los aspectos que caracterizan los textos literarios —estilo, creatividad, entorno, etc.— son complicados de procesar para los ordenadores; son factores que superan los límites de "comprensión" de los sistemas informáticos y suponen un gran reto para la traducción automática (Serrano, 2020).

Los primeros y tímidos intentos (Jones e Irvine, 2013; Bernardo, 2014) de emplear sistemas de traducción automática para textos literarios con Google Translate, un motor generalista, no dieron resultados satisfactorios: la traducción literal palabra por palabra no tenía en cuenta aspectos culturales ni contextuales, había demasiados errores léxicos, la calidad era muy baja y, en definitiva, se perdía la experiencia de lectura que sí proporciona una traducción humana.

Otros investigadores (Nunes Vieira, 2019) argumentan a favor de una buena utilización de la TA como herramienta de apoyo al traductor, y sugieren que los estudios de postedición de traducción literaria con TA dan resultados aceptables. Entre ellos encontramos los trabajos de Toral & Way (2015, 2018) que exploran una nueva perspectiva: entrenar motores de TA estadísticos y neuronales a partir de corpus paralelos y monolingües que sirvan específicamente para traducir literatura. Su primer experimento con un sistema estadístico para el par de lenguas español-catalán y la traducción de la obra de Carlos Ruiz Zafón *El prisionero del cielo* dio resultados bastante prometedores al compararla con una traducción humana: un 20% de las frases eran exactamente iguales a las propuestas por la traducción de referencia. Posteriormente, probaron con lenguas más alejadas como el inglés y el francés, y el resultado no fue tan

ventajoso, aunque la TA se sigue considerando como un buen punto de partida para la postedición.

Siguiendo con sus investigaciones Toral & Way (2018) introducen la comparación de un motor estadístico y otro neuronal, con resultados siempre favorables a este último y además, evalúan con traductores literarios profesionales el proceso de postedición de los resultados de la TA. Las conclusiones de este experimento indican claramente que, aunque no se puede negar un aumento considerable de la productividad (18% en el estadístico y 36% en el neuronal), los traductores no se sienten cómodos con este método de trabajo. Entre los aspectos más negativos destacan el de la división del texto en segmentos y, sobre todo, el de coartar su creatividad. Este aspecto de la creatividad es analizado y evaluado cuantitativamente en recientes trabajos de Guerberof & Toral (2021, 2022) y se llega a la conclusión de que la postedición constriñe la creatividad porque el traductor se ve más como un evaluador que como un creador; de ahí que los traductores prefieran la traducción desde cero. Estos autores proponen una definición de lo que se entiende por *creatividad* en traducción literaria:

"Creative in translation is the process of identifying and understanding a problem in the source text, generating several new and elegant solutions that depart from the source text and choosing the one that best fits the target text and culture to provide the reader the same experience as that of the source reader"(Guerberof & Toral, 2022, p.26).

Matusov (2019) ensaya esta misma perspectiva de adaptar un sistema de TA neuronal al contenido literario —con un corpus literario muy pequeño— para traducir historias de ficción del inglés al ruso y del alemán al inglés y lo compara con sistemas generalistas como Google Translate. Los resultados muestran que los sistemas adaptados tienen un léxico más rico y variado, el 30% de las oraciones tienen una calidad aceptable y hay pocos errores sintácticos incluso en oraciones largas, pero persisten los errores en palabras o expresiones ambiguas.

En conclusión, los últimos trabajos realizados en este campo apuntan a que, con los sistemas neuronales, cuyos *outputs* son el resultado de miles y miles de interconexiones de palabras y frases en miles y miles de contextos, podemos conseguir resultados que muchas veces son idénticos a los de la traducción humana y otras veces, sin ser iguales, tienen una calidad aceptable. En estas circunstancias, parece claro que los traductores

literarios contemplan la traducción automática como un desafío e incertidumbre para su profesión, pero no se cierran al apoyo técnico que le pueden proporcionar las herramientas tecnológicas. Desde esta perspectiva, los investigadores abogan por superar los temores y contemplan la integración de la tecnología en el trabajo del traductor de forma que le resulte de ayuda en su tarea (Youdale, 2020). Respecto a la TA, proponen la conveniencia de utilizar los resultados como sugerencias que le ayuden a decidir la mejor opción de entre las varias soluciones posibles y no solo como una propuesta de traducción que se tenga que posteditar (Toral & Way, 2015).

2.4 Evaluación de resultados de la TA

La evaluación de traducciones ha ido evolucionando a lo largo de los años, pero el objetivo siempre ha sido el de poder responder a la pregunta de si una traducción es buena o mala; es decir, establecer la calidad de la traducción. Juliane House (2012) realiza una revisión histórica de las distintas aproximaciones, desde las más subjetivas —mentalistas— que ponen el énfasis en las decisiones del traductor de las que depende la calidad, hasta otras más realistas que valoran la equivalencia directa y la inteligibilidad desde el punto de vista de los expertos y de los usuarios, o las teorías de los funcionalistas y *skopos* que ponen el foco en el propósito de la traducción y valoran su calidad por el grado de adaptación a las normas textuales y culturales de la lengua y cultura de llegada.

Por su parte, Hutchins (2010) señala tres aspectos que demuestran la calidad de una traducción: (i) la fidelidad al original, es decir, la precisión en que el texto traducido contiene la misma información (ii) la inteligibilidad o capacidad del lector para entender la traducción (iii) el estilo o en qué medida la traducción emplea un lenguaje apropiado a su contenido e intención.

Los estudios sobre evaluación de la calidad de las traducción dividen el campo de investigación en tres ámbitos (Saldanha & O'Brien, 2014): evaluar la calidad respecto al proceso de traducción, respecto al contexto y sobre todo, respecto al producto. Este último enfoque, el análisis de los resultados de traducción es, sin duda, el que más ha ocupado la atención de los expertos —lingüistas, traductores, investigadores—.

Una vez decidido el enfoque, es necesario tomar una serie de decisiones sobre el tipo de evaluación teniendo en cuenta una serie de variables, como la finalidad de la evaluación o el tipo de texto que se quiere evaluar. Las variables que definen nuestro proyecto y definirán el tipo de evaluación son: el tipo de texto —textos de ficción traducidos con TA— y la finalidad de la evaluación; es decir, comparar los resultados con los de otros sistemas de traducción automática.

Siguiendo las explicaciones de Antonio Toral[7], podemos distinguir tres tipos de evaluación de la TA: evaluación humana, evaluación automática y evaluación basada en tareas, la cual consiste en medir el esfuerzo y la transmisión de información.

En este trabajo, la evaluación no constituye el foco principal. Si bien se trata de una etapa que consideramos necesaria, no podemos abarcar todos los aspectos que se podrían tratar. Por tanto, nos hemos limitado a una evaluación automática de la traducción de algunos fragmentos de ficción literaria para comparar resultados en bruto, sin posteditar, de diferentes sistemas de TA.

2.4.1 Evaluación automática de sistemas de TA

En este tipo de evaluación, la premisa básica es comparar una frase o fragmento traducido con TA con esa misma frase o fragmento traducido por un traductor humano y que actúa como traducción de referencia. La métrica automática otorga al resultado de la TA una puntuación que representa la similaridad entre este resultado y la referencia humana considerada como correcta; de esta manera, cuanta mejor puntuación reciba una traducción, más calidad se le supone (Saldanha & O'Brien, 2014).

Las métricas automáticas no están pensadas para evaluar una frase individualmente, sin embargo, se ha demostrado empíricamente que si tenemos un conjunto de evaluación de muchas frases (500, 1000, 5000) la puntuación final correlaciona bastante con la que daría un evaluador humano, por tanto, son un indicador bastante fiable [6].

A continuación, definimos algunas métricas empleadas hoy en día según el estudio de Giménez (2008):

---

7 https://xwiki.recursos.uoc.edu/wiki/mat00001ca/view/Research%20on%20Translation%20Technologies/Automatic%20evaluation%20of%20Machine%20Translation/ Última consulta: 16 de mayo de 2022

● **BLEU**: (*Bilingual Evaluation Understudy*) Se trata de la métrica más popular y que representa el estándar de las mediciones automáticas. BLEU calcula cuántas secuencias de hasta 4 palabras en la salida del sistema coinciden con secuencias de hasta cuatro palabras en la referencia. Cuantas más coincidencias se hallen, más alta será su puntuación: 0 si nada coincide, 1 si todo coincide.

● **NIST**: esta métrica está basada en BLEU, aunque a diferencia de esta, otorga un mayor valor a las secuencias que son más informativas. De esta manera, NIST "pesa" la importancia de una secuencia según la frecuencia de aparición; en teoría, a menos apariciones, más información proporcionan. Al igual que BLEU, las puntuaciones elevadas determinan la buena calidad de la traducción.

● **WER**: (*Word Error Rate*) mide el índice de error por palabra. Mide el número mínimo de substituciones, supresiones o inserciones necesarias para una traducción válida. No proporciona detalles sobre la naturaleza de los errores de traducción. Un índice cercano a cero indica una traducción de buena calidad.

● **Distancia de edición**: Este algoritmo calcula el "coste" mínimo de transformar una cadena en otra por medio de la inserción, eliminación o reemplazo de caracteres.

● **TER**: (*Translation Error Rate*) evalúa las traducciones automáticas según el mínimo número de ediciones necesario para su correspondencia con la traducción de referencia. Los resultados son negativos, entendiéndose como valores óptimos aquellos más cercanos a cero.

Entre los aspectos más cuestionables de estas métricas automáticas, Saldanha & O'Brien (2014) citan la de otorgar a la traducción de referencia un papel de "traducción de oro" contra la que la candidata automática se tiene que comparar, cuando en realidad, no sabemos si la calidad de esta traducción humana de referencia ha sido debidamente evaluada. Creemos que este problema se acrecienta cuando se trata de traducciones de ficción literarias en las cuales, como venimos repitiendo, el factor de la creatividad del traductor puede jugar un papel muy importante. Aun así, se considera que las métricas automáticas son objetivas y no es desdeñable su validez cuando se trata de comparar dos o más sistemas de TA, tal y como demuestran los estudios de Toral & Way (2018) o Matusov (2019).

3. Metodología

En este capítulo se describe paso a paso la metodología MTUOC (apartado 1.3) empleada para la elaboración del corpus, el entrenamiento del motor de traducción automática neuronal y la evaluación de resultados.

Un buen ejemplo de aplicación del proyecto MTUOC es el TFM de Rodriguez del Rosario (2021) que sigue esta metodología para la creación de un corpus en el campo especializado de la aviación y el posterior montaje de motores de TA, uno estadístico y otro neuronal, para comparar sus resultados con los de sistemas genéricos, tanto estadísticos como neuronales.

En el repositorio online gratuito GitHub que gestiona Antoni Oliver[8] se alojan los programas para llevar a cabo las distintas partes de este estudio; es decir, los programas de descarga, segmentación y alineación del corpus paralelo; los programas de entrenamiento de sistemas neuronales; el servidor MTUOC que permite poner en marcha estos motores e integrarlos en entornos profesionales de traducción; el cliente de traducción MTUOC; y los programas para la evaluación de sistemas de traducción automática.

3.1 Objetivos e hipótesis

Los objetivos que queremos alcanzar se resumen en esta frase: Crear un corpus literario de gran tamaño para entrenar un motor de TA y comprobar si proporciona resultados de calidad comparables a los de otros sistemas de traducción generalista.

El proyecto de diseñar y elaborar un corpus de este tipo no está exento de retos y dificultades que, si no hemos podido superar, al menos hemos conseguido identificar, así como plantear posibles soluciones. Partimos de la hipótesis de que un motor de TA entrenado con contenido específico puede proporcionar mejores resultados que los de un sistema generalista en el contexto creativo literario. Comprobaremos mediante una evaluación con métricas automáticas si esta hipótesis se cumple.

[8] https://github.com/aoliverg Última consulta: 15 de mayo de 2022

3.2 Fases de elaboración del corpus

3.2.1 Diseño

Algunos de los trabajos revisados comparten con nosotros la elaboración de corpus especializados en literatura y su aplicación a la TA con distintas composiciones de sus corpus. Por un lado, Toral & Way (2018) para el par de lenguas catalán/inglés entrenan motores automáticos con una combinación de corpus paralelos y monolingües. Por otro lado, Matusov (2019) con el par de lenguas ruso/inglés entrena un motor adaptado a la literatura con un escaso corpus de obras literarias que combina con un motor generalista. En ambos casos, la comparación con los resultados de otros sistemas generalistas es favorable a los motores entrenados con corpus literario y sistema neuronal como hemos detallado en el apartado 2.3.1.

El tipo de corpus lingüístico en el que se centra este trabajo es un **corpus paralelo bilingüe** inglés-español de temática literaria elaborado con la finalidad de entrenar un motor de traducción automática. Los criterios seguidos para la selección de los textos vienen definidos por esta finalidad que marca las características que debe poseer nuestro corpus.

- En cuanto al tamaño: como hemos descrito en el apartado 2.2 y veremos más detenidamente en el 3.3, los sistemas de aprendizaje automático funcionan con una gran cantidad de datos para garantizar unos resultados fiables. Según A. Oliver (2020), alrededor de 5 millones de segmentos.

- En cuanto a la calidad de la alineación: debido al gran tamaño del corpus solo es viable realizar una alineación automática del mismo. En el apartado 3.2.4, describimos la utilización de un programa que nos permite seleccionar aquellos segmentos con una alineación más precisa.

- Respecto a la composición: el cuerpo central del corpus paralelo es de contenido literario. A él se suman, para completar el tamaño deseado, otros corpus de acceso público y de temática similar. El número de segmentos y las proporciones entre las diferentes partes se detallan en la Tabla 1:

Cuerpo central literario	Books[9] (Opus)	Corpus 100[10] (Opus)	QED-Educativo[11] (Opus)	Europarl[12] (Opus)	Global Voices (propio)	Corona Data[13](Taus)	**TOTAL**
1.334.526	88.601	855.438	1.052.575	1.040.582	50.850	897.425	**6.031.164**
22,12%	4.696.659 *77,87%*						*100%*

Tabla 1: Número de segmentos alineados EN-ES una vez completado el procesamiento

El 22,12% del corpus total es de contenido específicamente literario. En el punto 3.2.2 explicaremos el proceso de adquisición y procesamiento de los datos. Los restantes corpus se han recogido en fuentes de acceso libre, concretamente de los corpus recopilados en la web *Opus Corpus* (Tiedemann, 2012) y de las contribuciones de *TAUS* y de Global Voices. Los corpus de Opus no han necesitado ulterior procesamiento puesto que se pueden descargar en el formato alineado que precisamos; en cambio, el *Corona Corpora* de *TAUS* y el corpus de *Global Voices* fueron descargados y procesados en el transcurso de las prácticas de investigación que realizamos el curso pasado (periodo 2020-2021). Todos ellos se han unido para formar un gran corpus general del cual, como explicaremos en el apartado 3.3.1, se escogen los segmentos más similares a los del corpus literario.

3.2.2 Adquisición y preparación del corpus paralelo

El corpus central está compuesto de obras literarias en formato electrónico la mayoría de ellas convertidas a texto desde el formato epub. Las fuentes de las que provienen son sitios web de acceso público que permiten la descarga libre de libros con los derechos de autor caducados. El grueso de los materiales procede de la web *Project Gutenberg*[14], que contiene una colección de más de 65.000 libros en formato epub, la mayoría de ellos en inglés, pero también en otros idiomas. Otra fuentes para obras en español han

[9] https://opus.nlpl.eu/Books.php Última consulta: 4 de abril de 2022
[10] https://opus.nlpl.eu/opus-100.php Última consulta: 4 de abril de 2022
[11] https://opus.nlpl.eu/QED.php Última consulta: 4 de abril de 2022
[12] https://historicalarchives.europarl.europa.eu/home.htm Última consulta: mayo de 2021
[13] https://md.taus.net/corona Última consulta: mayo de 2021
[14] https://www.gutenberg.org/ Última consulta: abril de 2021

sido la Biblioteca Nacional[15] y la Biblioteca Calibre[16], que pueden ser descargadas para seleccionar aquellas obras que están traducidas a los dos idiomas.

Para la descarga y procesamiento de los archivos del sitio *Project Gutenberg,* que se llevó a cabo durante los meses febrero y marzo de 2022, se siguieron las siguientes fases, según la propuesta y los programas elaborados por Antoni Oliver:[17]

1. El primer paso fue instalar wget[18], una herramienta gratuita para la descarga de archivos a través de la línea de comandos y que permite, por ejemplo, reanudar una descarga interrumpida. Esta aplicación funciona en el entorno del sistema operativo Linux, o en su defecto, con WSL (*Windows Subsystem for Linux*), que es el entorno que utilizamos en este proyecto para ejecutar los programas.

2. Con un sencillo programa de Python, *createscript.py*, se creó un *script SH que* contiene una instrucción *wget* para descargar los epub de manera secuencial y continua, y un *sleep* aleatorio que hace que se espere un poco a bajar el siguiente y no se detecte como bajada masiva. La descarga duró casi tres días.

3. Una vez descargados (63.852 epubs), ejecutamos el programa *epub2CatalogDIR.py* para la creación de un catálogo general con los nombres de las obras descargadas, desde el cual seleccionamos aquellas obras que nos interesaban por tener la traducción correspondiente entre el par de lenguas inglés-español. El catálogo de obras es muy extenso, pero traducciones no hay tantas.

4. La selección de obras con su correspondiente traducción se realizó manualmente y dio como resultado un listado de 384 obras traducidas para el par inglés/español. El listado de obras antes de su procesamiento se puede ver en el Anexo I.

La descarga de archivos de la Biblioteca Calibre se realizó a través de un *magnet link* que permite el acceso directo a la descarga de contenidos a través del programa BitTorrent, mientras que los libros de la Biblioteca Nacional seleccionados fueron descargados uno a uno.

[15] http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/ Última consulta: abril de 2021
[16] https://laleyendadesw.mforos.com/1493088/11588310-biblioteca-calibre-30026-ebooks-epubs/ Última consulta: abril de 2021
[17] https://xwiki.recursos.uoc.edu/wiki/matm21564es/view/test Última consulta: abril de 2021
[18] https://www.gnu.org/software/wget/ Última consulta: abril de 2021

Como hemos indicado anteriormente (apartado 2.1.1), los archivos para crear el corpus debían estar en formato de texto para ser procesados en las siguientes fases de elaboración del corpus. La gran mayoría de obras disponibles estaban en formato epub; por tanto, la primera tarea fue convertirlos a texto con el programa *epub2txt.py*[19]. Al mismo tiempo, creamos dos directorios (uno para cada idioma), en el que fuimos incorporando paralelamente las obras, original y traducida, una vez convertidas a formato texto y que debían tener el mismo nombre con la única diferencia del sufijo -EN o -ES respectivamente. Esto es importante para la fase de alineación posterior. Una vez tuvimos los archivos ordenados en sus correspondientes directorios, procedimos a la segmentación y alineación automática de los textos.

3.2.3 Segmentación

El proyecto MTUOC facilita una serie de programas de utilidades[20] para procesar los archivos y prepararlos para la creación del corpus. El primer paso es la segmentación, es decir, "dividir los párrafos en unidades más pequeñas del estilo de la oración" (A. Oliver, 2021) utilizando el programa *txt2segmentedtextDIR.py,* que nos permite segmentar todos los archivos del directorio a la vez. El programa utiliza un archivo estándar de reglas de segmentación para varias lenguas —*segment.srx*— y un programa segmentador —*srx.segmenter.py*—. Además, añade la marca <*p*> de separación entre segmentos. Esta marca servirá para la fase posterior de alineación con *Hunalign.*

3.2.4 Alineación

Existen programas de alineación semi-automática como *bitext2tmx* o *LF Aligner* que son muy útiles para alinear textos cortos y que permiten una revisión manual segmento por segmento, pero cuando se trata de directorios con cientos de archivos y miles de segmentos cada uno, es necesario utilizar un programa de alineación automática.

Hunalign (Varga et al., 2005) es un programa que alinea texto bilingüe a nivel oracional de manera automática. Para su ejecución, requiere texto tokenizado y segmentado por

[19] https://pypi.org/project/EbookLib/ Última consulta: mayo de 2022
[20] https://github.com/aoliverg/MTUOC-utils Última consulta: 14 de mayo de 2022

oraciones en dos idiomas. En el capítulo dedicado a la alineación[21] entre los recursos del máster de Traducción y Tecnologías de la UOC, encontramos detallados los pasos que se han seguido. Además de los textos segmentados, Hunalign utiliza un diccionario bilingüe[22] que le ayuda en la alineación. Si no se le proporciona el diccionario bilingüe, Hunalign lo creará basándose en la información de la longitud de las oraciones para después volver a alinear el texto con este diccionario automático. La salida del programa, en el caso más simple, es una secuencia de pares de oraciones bilingües. Como la mayoría de los programas de alineación automática, Hunalign no se ocupa de los cambios en el orden de las oraciones; es decir, los segmentos A y B en una lengua se alinean con los segmentos A y B en la otra lengua, aunque la correspondencia de significados implique un cambio de orden[23].

Hunalign permite la alineación de todos los archivos de texto que estén organizados en dos directorios paralelos, uno para cada lengua, tokenizados, segmentados y con la etiqueta <p> de separación de segmentos. Para ello, ejecutamos el programa hunalign en *batch mode* o trabajo por lotes. Estos son los pasos:

1. Creamos un *script de alineación* (Ilustración 1) que contiene el conjunto de órdenes para realizar una acción; en este caso, para que alinee todos los archivos de los directorios para cada lengua:

```
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AgathaAffair-EN.txt" "textos-seg-ES/AgathaAffair-ES.txt" > "TEXTOS-ALI//ali-AgathaAffair-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AgathaBrown-EN.txt" "textos-seg-ES/AgathaBrown-ES.txt" > "TEXTOS-ALI//ali-AgathaBrown-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AgathaMurder-EN.txt" "textos-seg-ES/AgathaMurder-ES.txt" > "TEXTOS-ALI//ali-AgathaMurder-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AgathaPoirot-EN.txt" "textos-seg-ES/AgathaPoirot-ES.txt" > "TEXTOS-ALI//ali-AgathaPoirot-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AnatoleAngels-EN.txt" "textos-seg-ES/AnatoleAngels-ES.txt" > "TEXTOS-ALI//ali-AnatoleAngels-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AnatoleCrime-EN.txt" "textos-seg-ES/AnatoleCrime-ES.txt" > "TEXTOS-ALI//ali-AnatoleCrime-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AnatoleJudea-EN.txt" "textos-seg-ES/AnatoleJudea-ES.txt" > "TEXTOS-ALI//ali-AnatoleJudea-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AnatoleRing-EN.txt" "textos-seg-ES/AnatoleRing-ES.txt" > "TEXTOS-ALI//ali-AnatoleRing-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AnatoleThais-EN.txt" "textos-seg-ES/AnatoleThais-ES.txt" > "TEXTOS-ALI//ali-AnatoleThais-EN.txt"
./hunalign hunapertium-en-es.dic -utf -realign -text "textos-seg-EN/AristotPolitics-EN.txt" "textos-seg-ES/AristotPolitics-ES.txt" > "TEXTOS-ALI//ali-AristotPolitics-EN.txt"
```

*Ilustración 1: Fragmento del script de alineación con las instrucciones para cada uno de los archivos. Se trata de un archivo **batch** que permite el trabajo por lotes, en orden secuencial y sin parar.*

21 http://xwiki.recursos.uoc.edu/wiki/matm4957ca/view/Automatic%20text%20alignment%20with%20Hunalign/ Última consulta: mayo de 2022

22 https://github.com/aoliverg/hunapertium Última consulta: 10 mayo de 2022

23 https://github.com/danielvarga/hunalign Última consulta 30 de mayo de 2022

2. Una vez ejecutado el programa *hunalign*, obtuvimos un directorio de archivos que concatenamos — *cat * > ../alineacio-unica-en-es.txt*— a un único archivo con segmentos alineados en los dos idiomas con la marca de párrafo <p> y con un índice de fiabilidad. En general, las alineaciones con un índice 0 o superior a 0 se consideran buenas; si el índice es negativo, son malas. En la ilustración 2 vemos algunas alineaciones de ejemplo:

Segmento original	**Segmento traducido**	**Puntuación de fiabilidad**
"What's wrong?	¿Qué marcha mal?	0.278571
"You do believe then?" he exclaimed	—¿Entonces me cree? —exclamó.	0.244444
I felt his body grow tense	Me di cuenta de que su cuerpo se tensaba	0.156522
"Now," he told me.	—Ahora —me dijo—, no tengo miedo.	0.05625
"I do not fear.	Si yo…	-0.05
"Use it on this—and I will show you."	Utilizadlo sobre esto… y os mostraré algo.	-0.0387097
She began eagerly	Comenzó a hablar con nerviosismo	-0.0375
<p>	<p>	0

Ilustración 2: Segmentos alineados con puntuación de fiabilidad. La marca de separación <p> aparece después de cada segmento o párrafo.

En una primera prueba, nos dimos cuenta de que había errores de segmentación y alineación y, aunque el propósito del corpus no exigía que este fuera perfecto (Zanettin, 2014), pensamos que sería conveniente retocar manualmente los textos que contenían más problemas con ayuda del editor de textos *Notepad++*, e incluso eliminamos algunos archivos cuando la segmentación era errónea. No obstante, pronto comprobamos que la tarea de editar manualmente 384 libros por cada idioma nos llevaría un tiempo excesivo. Por tanto, confiamos en la selección que proporciona el programa de alineación automática *Hunalign* que es capaz de reconocer los segmentos "no fiables" y los puede eliminar automáticamente como veremos a continuación.

4. Ejecutamos el programa *selectAlignements.py* para eliminar la etiqueta <p>, así como todos los segmentos cuyo índice de fiabilidad fuera menor que 0. El archivo final fue un corpus paralelo alineado de 1.334.526 segmentos. Después de comprobar que

persistían problemas en la alineación, decidimos seguir probando índices de fiabilidad más altos cuyos resultados detallaremos en el capítulo 4.

3.3. Creación de un motor de TA

Los sistemas de traducción automática están basados en algoritmos del *aprendizaje automático*[24] que pueden aprender y hacer predicciones sobre los datos. Con el fin de hacer estas predicciones o decisiones, los algoritmos construyen un modelo matemático a partir de los datos de entrada. El entrenamiento del motor de TA consiste precisamente en la construcción de ese modelo de aprendizaje a partir de unos datos —el corpus paralelo—que requieren una preparación antes de iniciar el entrenamiento.

3.3.1 Preparación: Limpieza, combinación y pre-procesamiento del corpus

Explicaremos brevemente en qué ha consistido esta preparación del corpus que es prácticamente la misma para todos los sistemas de traducción automática, ya sean estadísticos o neuronales.

En el apartado 3.2.1 se han detallado los distintos corpus que hemos concatenado en un único archivo para formar el corpus paralelo general que, posteriormente, combinamos con el nuestro. Todos los corpus —general y literario— han sido sometidos al **programa de limpieza**[25] que realiza, por defecto, las siguientes acciones: normalizar el apóstrofe, eliminar etiquetas, deshacer entidades HTML, quitar segmentos alineados iguales, eliminar segmentos con más del 60% de números y eliminar segmentos de menos de 5 caracteres.

El **programa de combinación**[26] funciona de la siguiente manera: el *input* que proporcionamos está formado por nuestro corpus de tamaño mediano (1,3 millones de segmentos) de tema específico literario, junto con el corpus de tamaño grande (4,7 millones). El programa selecciona aquellos segmentos del corpus general que sean lo más parecidos posible al corpus específico siguiendo un modelo de lengua creado a

[24] https://hmong.es/wiki/Training_set Última consulta: 22 de mayo de 2022
[25] https://github.com/aoliverg/MTUOC-clean-parallel-corpus Última consulta: 5 de abril de 2022
[26] https://github.com/aoliverg/MTUOC-corpus-combination Última consulta: 5 de abril de 2022

partir de los segmentos de la lengua de partida. Utiliza una base de datos *SQLite* para almacenar todos los datos, por lo que es capaz de trabajar con corpus extraordinariamente grandes. El *output* del programa resultó en un corpus de unos 6 millones de segmentos que conforman los datos de entrada para construir el modelo de aprendizaje final, es decir, para entrenar nuestro motor de TA.

En las diferentes etapas de la creación del modelo, se utilizan distintos conjuntos de datos; por esta razón, el programa de combinación divide el corpus en tres subconjuntos, como vemos en la ilustración 3:

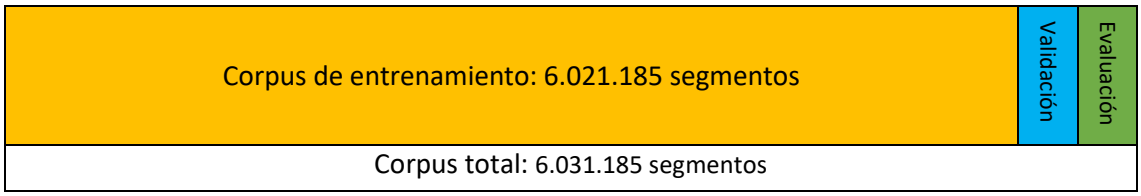

Ilustración 3: Partición del corpus en tres subconjuntos. Los corpus de validación y evaluación constan de 5000 segmentos cada uno.

- **Corpus de entrenamiento —train:** es el fragmento que contiene la mayor parte del total de los segmentos y que servirá para entrenar el sistema; es decir, para conseguir un modelo ajustado (entrenado) que sea capaz de predecir datos nuevos y desconocidos (traducciones).
- **Corpus de validación—val:** los 5000 segmentos —es la cifra habitual— que lo conforman proceden en su totalidad del núcleo del corpus, es decir, de nuestro corpus literario, y no han servido para entrenar el sistema. Un *conjunto de datos de validación* es un conjunto de datos de ejemplos que se utilizan para ajustar la arquitectura del sistema, y optimizarlo en cada iteración de entrenamiento. El corpus de validación proporciona una evaluación imparcial de los ajustes que se van produciendo durante el entrenamiento y permiten estimar la precisión y sensibilidad del modelo[27].
- **Corpus de evaluación—eval:** es el conjunto de 5000 segmentos que se reservan para la evaluación del sistema tanto en la lengua original —texto de partida—, como en la lengua de llegada —el subconjunto que actúa como traducción de referencia.

[27] https://hmong.es/wiki/Training_set Última consulta: 23 de mayo de 2022

El último paso de la preparación es la ejecución del **programa de preprocesamiento**[28] del corpus que realiza la siguiente serie de acciones:

- *Tokenización* del corpus tanto de la lengua de partida como de llegada. El proceso de *tokenización* consiste en dividir todas las unidades léxicas en *tokens;* es decir, separar las palabras, los signos de puntuación y las cifras con espacios.
- T*ruecasing*. Un *truecaser* es un programa que convierte todas las palabras (*tokens*) de un segmento en su forma básica —mayúscula o minúscula— independientemente de la posición que ocupen en la frase. Se entrena un truecaser para la lengua de partida y uno para la de llegada utilizando el corpus y, opcionalmente, un diccionario de palabras con todas sus formas flexionadas. Una vez entrenado el *truecaser* se aplica al corpus de entrenamiento.
- Reemplazo de emails, de URLs y de expresiones numéricas por un código (por defecto @EMAIL@, @URL@, @NUM@) que luego se podrán recuperar en la traducción definitiva.
- La técnica de procesamiento utilizada ha sido *Sentence Piece*. Antes de iniciar el entrenamiento es necesario delimitar la extensión del vocabulario —el número de palabras— sobre las que trabajará el sistema. *SentencePiece* genera un vocabulario implementando una estrategia denominada *subword* —fragmento de palabra— junto con un modelo de lenguaje *unigram* usando el corpus de entrenamiento de las lenguas en juego. De esta manera se pueden representar todas las palabras, ya sean enteras o por fragmentos: las palabras más frecuentes (formas) se utilizan enteras, pero las menos frecuentes se dividen en trozos que sean frecuentes. El programa *SentencePiece* se puede utilizar como algoritmo para calcular y aplicar sub-palabras sobre un corpus ya tokenizado y *truecased*, así como preprocesador del corpus sin necesidad de los pasos previos. La elección del tipo de preprocesamiento depende de los textos a traducir: en textos muy libres como los literarios donde pueden aparecer, por ejemplo,

[28] https://github.com/aoliverg/MTUOC-corpus-preprocessing Última consulta: 5 de abril de 2022

muchos fragmentos en mayúsculas, es recomendable realizar los pasos por separado[29].

- Por último, el *Guided Alignment*[30] o alineación a nivel de palabra —o sub-palabra— se utiliza para entrenar sistemas neuronales de tipo *transformer* (cap.2.2) que necesitan esta información para que la oración traducida con TA reproduzca el alineamiento a nivel de palabra o sub-palabra de la oración de partida. En este trabajo, hemos elegido el programa *fast_align*[31] para calcular la alineación a nivel de palabra (o sub-palabra) del corpus de entrenamiento y validación.

Al final del proceso se generaron los archivos que se utilizaron para entrenar y montar un sistema de TA neuronal tipo transformer. Los archivos son los siguientes:

Del corpus de entrenamiento: **train.sp.en** y **train.sp.es**

Del corpus de validación: **val.sp.en. en** y **val.sp.es**

Guided Alignment nos proporciona**: train.sp.en.es. align** y **val.sp.en.es. align**

3.3.2 Entrenamiento e instalación del sistema

Marian es un software de entrenamiento de acceso libre desarrollado principalmente por el equipo de Microsoft Translator para entrenar motores de traducción automática. Los programas que lo componen se pueden descargar libremente desde el sitio de GitHub[32], pero requieren un conocimiento informático muy especializado. Gracias al proyecto MTUOC, que ofrece unos scripts simplificados, se puede llevar a cabo el procedimiento de manera más accesible y menos complicada. El **proceso de entrenamiento**[33] requiere un hardware potente con unidades de procesamiento gráfico (GPU) para generar el modelo de aprendizaje que servirá para traducir. Además, durante el entrenamiento, se crean multitud de archivos temporales que sobrecargan una

[29] https://github.com/aoliverg/a_practical_course_on_machine_translation/wiki/C1.-Pasos-gen%C3%A9ricos-de-procesamiento Última consulta: 23 de mayo de 2022
[30] https://github.com/aoliverg/a_practical_course_on_machine_translation/wiki/C2.-Script-de-preprocesamiento Última consulta 15 de mayo de 2022
[31] https://github.com/clab/fast_align Última consulta: 10 de mayo de 2022
[32] https://marian-nmt.github.io/ Última consulta: 14 de mayo de 2022
[33] https://github.com/aoliverg/a_practical_course_on_machine_translation/wiki/C3.-Entrenamiento-con-Marian Última consulta: 10 de mayo de 2022

unidad que no esté equipada con suficiente memoria RAM. Al no disponer de este tipo de unidades, solicitamos y obtuvimos la ayuda del profesor Antoni Oliver que realizó el entrenamiento del sistema con nuestros archivos pre-procesados. Por nuestra parte, pusimos en marcha el entrenamiento durante unos minutos para comprobar su funcionamiento.

La posterior **instalación** del motor de TA consiste en agrupar en un mismo directorio los siguientes archivos y programas:

- Los archivos generados durante el entrenamiento.

- El archivo compilado *marian-server*.

- Los programas que componen el servidor de MTUOC.

3.3.3 Puesta en marcha y traducción con MTUOC

Una vez instalado, el motor se comporta como un servidor que, al poner en marcha mediante la ejecución del programa *MTUOC-server.py*, nos facilita una *dirección IP*, un *puerto de conexión* y un *servidor de traduc*ción que utilizaremos para traducir.

El **servidor MTUOC**[34] es compatible con otros servidores de traducción —Marian, OpenNMT, Moses o ModernMT— y, por tanto, facilita la integración de los sistemas de traducción automática en el flujo de trabajo de distintas herramientas de traducción asistida como *OmegaT*, *Okapi o SDL Trados.*

El **programa-cliente MTUOC-Translator**[35] es otra de las aplicaciones del proyecto que hemos utilizado en este trabajo y funciona de la siguiente manera (ilustración 4): el cliente envía una petición de traducción de un texto que debe estar segmentado; el servidor MTUOC preprocesa el texto—tokenización, SentencePiece, etc.— y establece una comunicación con uno de los servidores de traducción compatibles, mediante la dirección IP y el puerto proporcionados al ponerse en marcha. El servidor de traducción, que puede ser MTUOC-server, Marian, OpenNMT, Moses o ModernMT, devuelve la traducción para que el servidor de MTUOC la desprocese antes de devolverla al cliente MTUOC-Translator como un segmento correcto en la lengua de llegada.

[34] https://github.com/aoliverg/MTUOC-server Última consulta: 14 de mayo de 2022
[35] https://github.com/aoliverg/MTUOC-translator Última consulta: 14 de mayo de 2022

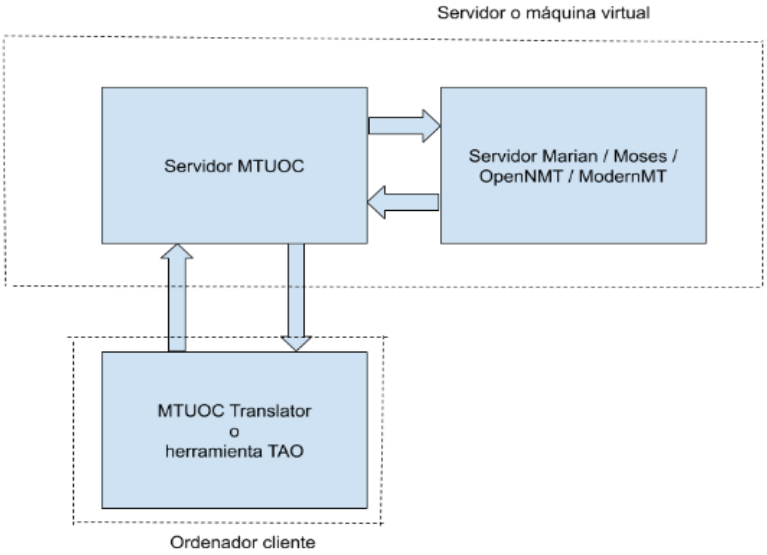

Ilustración 4: Funcionamiento del servidor MTUOC. Imagen extraída de XWiki del Máster de Traducción y Tecnologías

Con el corpus de evaluación —**eval**— que hemos reservado del corpus y que, recordemos, no se utiliza para entrenar el motor, hemos probado el funcionamiento del Traductor de MTUOC solicitando una traducción del archivo en inglés al español, para después someter los resultados de traducción a una evaluación.

4. Resultados

4.1 Relativos al proceso de creación del sistema

Los primeros resultados de traducción, incluso antes de la evaluación formal, ya nos mostraron algunos errores en el sistema que habíamos entrenado: mientras que la mayoría de frases traducidas eran aceptables, aparecían segmentos sin sentido como los que vemos en la ilustración 5, que se repetían cada cierto intervalo de tiempo:

— ¡Oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡

— Sí, sí, sí, sí, sí, sí, sí, sí, sí, sí, sí.

— ¡Oh!, ¡ah!, ¡ah!, ¡ay!, ¡ay!, ¡ay!, ¡ay!, ¡ay!, ¡ay!, ¡ay!, ¡ay!, ¡ay!

Ilustración 5: Ejemplos de outputs sin sentido.

Al investigar donde podría estar el problema, nos dimos cuenta de que el corpus de validación contenía problemas de alineación (Anexo II), con lo que los valores de validación no eran suficientemente altos y, aunque no llegaron a forzar la parada, dieron lugar a este tipo de errores en la traducción.

Con el propósito de mejorar los resultados se aplicaron dos posibles soluciones:

1ª) Mejorar la segmentación y alineación del corpus y probar otro entrenamiento.

Se llevó a cabo una nueva segmentación de los textos en las dos lenguas y procedimos a probar distintas alineaciones automáticas subiendo el índice de fiabilidad para seleccionar aquellos segmentos con más alta precisión. En la tabla 2 podemos comprobar cómo subir el índice de fiabilidad supone perder muchos segmentos:

Núm. de segmentos	Índices de fiabilidad
1.435.160	Alineación 0
1.153.782	Alineación 0.3
540.298	Alineación 0.5
471.788	Alineación 0.6
287.895	Alineación 0.8
159.840	Alineación 1.0

Tabla 2: El número de segmentos del corpus disminuye a medida que se incrementa el índice de fiabilidad

En esta coyuntura, observamos que la solución aplicada puede proporcionar segmentos bien alineados, pero, en realidad, muchos de estos segmentos no aportan nada como vemos en la ilustración 6 y, además, hay una pérdida importante de información válida que una alineación manual hubiera mantenido:

‘No, sir.' ----------	—No, señor.
‘No, sir.' ----------	—No, señor.
‘No, sir.' ----------	—No, señor.
“Yes, sir. ----------	—No, señor.
“Yes, sir. ----------	—No, señor.
“No, sir. ----------	—No, señor.
CHAPTER XV ----------	CAPÍTULO XV
CHAPTER XVI ----------	CAPÍTULO XVI
CHAPTER V ----------	CAPÍTULO V
CHAPTER VI ----------	CAPÍTULO VI
CHAPTER VII ----------	CAPÍTULO VII
CHAPTER VIII ----------	CAPÍTULO VIII
CHAPTER IX ----------	CAPÍTULO IX
CHAPTER X ----------	CAPÍTULO X
CHAPTER XI ----------	CAPÍTULO XI

Ilustración 6: Ejemplos de segmentos que quedan tras un índice de fiabilidad de 2.0. El corpus final contendría 8.117 segmentos, una ínfima parte de los 1.334.526 del corpus con alineación 0.0

Hemos puesto en marcha dos nuevos pre-procesamientos con los corpus resultantes de la aplicación de los índices 0.3 y 0.8 y en ambos casos hemos obtenido corpus de validación con alineaciones más precisas, pero durante el entrenamiento los valores de validación seguían siendo mínimos y no proporcionaban mejores resultados. Ante esta situación, decidimos detener los entrenamientos. En el archivo valid.log (Anexo II) se detallan y analizan todos los eventos que ocurren durante el entrenamiento. Podemos observar que el entrenamiento no avanza o lo hace demasiado lentamente.

Un análisis más detallado de los resultados de las distintas alineaciones nos indica que, incluso siendo pocos los segmentos seleccionados, persisten los problemas de alineación y que no siempre los segmentos descartados son malos. En las conclusiones detallaremos las posibles causas de estas anomalías.

A pesar de estos problemas, el sistema funciona correctamente y establece las conexiones pertinentes con los servidores de traducción (ver apartado 3.3.3). Hemos llevado a cabo varias traducciones de diferentes textos gracias a que el corpus de entrenamiento está formado por la combinación de todos los componentes del corpus final (ver apartado 3.2.1), el cual contiene una mayoría de segmentos bien alineados.

2) La segunda solución posible era probar un set de evaluación con textos bien segmentados y alineados, puesto que el corpus reservado para evaluación también contenía errores. En concreto, preparamos la novela *The curious incident of the dog in*

the night-time de Mark Haddon de manera automática, pero en solitario, con un resultado de 3500 segmentos alineados (ver fragmento en el Anexo III).

En esta ocasión, los resultados de traducción fueron aceptables, sin segmentos absurdos y perfectamente comparables a los de otros sistemas y métodos de traducción como podemos observar en los ejemplos de la tabla 3:

Tabla 3: Comparación de traducciones

Texto original	**Traducción de referencia**	**Traducción Google Translate**	**Traducción MTUOC**
But the dog was not running or asleep.	Pero el perro no estaba corriendo o dormido.	Pero el perro no estaba corriendo ni dormido.	Pero el perro no se movía ni dormía.
The policeman had a big orange leaf stuck to the bottom of his shoe which was poking out from one side.	El policía llevaba pegada a la suela del zapato una gran hoja naranja, que le sobresalía por un lado.	El policía tenía una gran hoja de naranja pegada al fondo de su zapato que estaba empujando de un lado.	El policía tenía una gran hoja de naranja pegado a la parte inferior de su zapato, que salía de un lado.
“And what, precisely, were you doing in the garden?” he asked. “I was holding the dog,” I replied. “And why were you holding the dog?” he asked.	— ¿Y qué hacías exactamente en el jardín? — preguntó. — Tenía al perro en brazos —dije. — ¿Y por qué tenías al perro en brazos? — preguntó.	"¿Y qué, precisamente, estabas haciendo en el jardín?" preguntó. "Estaba sosteniendo al perro", respondí. "¿Y por qué estabas sosteniendo al perro?" preguntó.	— ¿Y qué hacías en el jardín? — preguntó. — Yo estaba sujetando al perro — repliqué —. — ¿Y por qué estaba usted con el perro? — preguntó.
These are examples of metaphors: I laughed my socks off. He was the apple of her eye. They had a skeleton in the cupboard. We had a real pig of a day.	He aquí ejemplos de metáforas: Me he reído mucho. Era la niña de sus ojos Tenían un cadáver en el armario. Pasamos un día de mil demonios.	Estos son ejemplos de metáforas: Me reí mis calcetines. Él era la manzana de su ojo. Tenían un esqueleto en el armario. Tuvimos un verdadero cerdo de día.	Estos son ejemplos de metáforas: Me reí mis calcetines. Era la manzana de sus ojos. Tenía un esqueleto en el armario. Teníamos un verdadero cerdo de un día.
En rojo: errores de traducción; En verde: aciertos; En azul: falta de concordancia o adecuación			

Posteriormente, seleccionamos manualmente 500 segmentos con el programa de alineación LF Aligner[36] para comprobar si, efectivamente, una precisión mayor de la alineación entre el texto original y el texto de referencia influye en la puntuación que las métricas automáticas otorgan a las traducciones.

Por último, hemos traducido un capítulo del *Hobbit* de Tolkien (ver fragmento en Anexo IV) para seguir comprobando el rendimiento del motor utilizando la herramienta *Rainbow* de *Okapi Framework*[37] que conectó con el servidor de traducción ModernMT.

4.2 Resultados relativos al producto final: evaluación con métricas automáticas

El tercer objetivo de este trabajo era evaluar los resultados de traducción proporcionados por nuestro servidor de traducción y compararlos con de un sistema de TA genérico y potente como Google Translate. Con ello, podríamos corroborar o rebatir nuestra hipótesis de que un motor preparado específicamente con contenido literario puede proporcionar mejores resultados que los de un sistema generalista.

El procedimiento de evaluación con métricas automáticas consiste en otorgar una puntuación a la traducción que queremos evaluar sobre la base de una traducción de referencia, generalmente, hecha por un traductor humano. Este mismo procedimiento se ha aplicado a la traducción obtenida en Google Translate. Para aplicar el **programa de evaluación**[38], se requiere texto segmentado y un tokenizador del idioma que se evalúa. El traductor de MTUOC devuelve el texto segmentado tal y como lo introducimos; y para obtener texto segmentado de Google Translate usamos *Google Sheets* mediante la fórmula:

=GOOGLETRANSLATE (A1, “en”,” es”)

que permite conectar directamente con el traductor de Google manteniendo la segmentación del texto. Resumimos la lectura de las métricas utilizadas (ver apartado 2.4.1) que proporciona el programa MTUOC-eval [39]:

[36] https://sourceforge.net/projects/aligner/ Última consulta: 20 de mayo de 2022
[37] https://okapiframework.org/ Última consulta: 30 de mayo de 2022
[38] https://github.com/aoliverg/MTUOC-eval Última consulta: 15 de mayo de 2022
[39] https://github.com/aoliverg/MTUOC-eval/wiki Última consulta: 15 de mayo de 2022

BLEU: Mide las coincidencias entre secuencias (referencia-traducción) siendo 0 la mínima puntuación y 1 la máxima que puede otorgar.

NIST: Cuánto más alta sea la puntuación, más coincidencias relevantes encuentra con la referencia.

WER: Un índice de error por palabra cercano a 0 indica que se han efectuado un mínimo de correcciones necesarias para una traducción válida.

%EdDist: El coste de edición mínimo necesario para transformar una cadena en otra. A menos coste, más semejanza con la referencia.

TER: Calcula el mínimo número de ediciones necesario para su correspondencia con la traducción de referencia. Los valores óptimos son los más cercanos a 0.

En primer lugar, se realizaron dos evaluaciones distintas en un intervalo de 15 días, para cotejar los resultados de los dos conjuntos de evaluación. En la tabla 4 podemos comparar las puntuaciones de las distintas métricas automáticas:

Tabla 4: Comparativa de evaluaciones de 3500 segmentos y 500 segmentos del corpus eval.

	Primera evaluación: 3500 K			Segunda evaluación: 500 K	
	Google Translate	**MTUOC**		**Google Translate**	**MTUOC**
BLEU	0.2903	0.2948	BLEU	0.3548	0.3544
NIST	6.383	6.389	NIST	6.528	6.590
WER	0.7641	0.8068	WER	0.5991	0.6169
%EdDist	49.08	49.66	%EdDist	40.40	40.33
TER	0.6479	0.6351	TER	0.5346	0.5121

En ambas evaluaciones, la comparación de las traducciones entre MTUOC y Google Translate da unos resultados muy similares. Esto nos indica que nuestro motor es capaz de producir traducciones de una calidad comparable a la del motor generalista.

La comparación entre las evaluaciones de 3500K y 500K muestra una mejoría de la segunda respecto a la primera. Con estos datos se puede argumentar que la alineación en el texto de referencia también afecta a los resultados de la evaluación; al seleccionar un conjunto de evaluación con una alineación más precisa, obtenemos mejores puntuaciones. Como ya hemos descrito en el apartado 2.4.1, uno de los problemas de las métricas automáticas es el de "suponer" que la traducción de referencia es una traducción ideal; si este aspecto puede ser problemático en traducciones de temática especializada, en traducción literaria es casi imposible definir cómo sería esta "traducción ideal". Esto implica que las métricas automáticas ofrecen puntuaciones que pueden ser válidas en términos comparativos entre distintos sistemas de traducción,

pero, al menos en el ámbito de la literatura, no pueden valorar objetivamente que un texto tenga mejor o peor calidad.

En segundo lugar, y para terminar este capítulo de evaluaciones, en la tabla 5 vemos los resultados de la evaluación del fragmento del *Hobbit,* para cuya traducción nuestro servidor contactó con ModernMT a través de Rainbow de Okapi:

	Evaluación capítulo 6 del Hobbit	
	Google Translate	**MTUOC-Rainbow**
BLEU	0.2483	0.2209
NIST	5.812	5.100
WER	0.6297	0.7697
%EdDist	49.15	52.53
TER	0.5618	0.6284

Tabla 5: Comparación de puntuaciones de distintas métricas para dos sistemas de traducción

Podemos observar una ligera mejor puntuación de la traducción de Google Translate. Este resultado confirma que persisten los problemas con nuestro sistema y nos reafirma en la idea de que existe un margen real de mejora si se solucionan los errores de alineación del corpus.

5. Conclusiones

5.1 Contribución al campo de estudio

El propósito de este trabajo era crear un corpus paralelo literario para aplicarlo al montaje y puesta en marcha de un motor de TA y evaluar comparativamente la calidad de sus resultados y los resultados de Google Translate con métricas automáticas objetivas. Con ello pretendíamos tender un puente entre dos modalidades de traducción muy distintas y distantes como son la traducción automática y la literaria. Después de un breve recorrido por la literatura en torno a la creación de corpus paralelos, los sistemas de TA y la traducción literaria realizamos una revisión del contexto en el que se desarrolla actualmente la traducción automática de textos literarios siguiendo, sobre todo, las investigaciones de Toral (2015,2018) y Matusov (2019) que experimentan con la creación de distintos motores de TA y comparan sus resultados con otros sistemas generalistas.

Para llevar a cabo este estudio hemos utilizado las herramientas y la metodología del proyecto MTUOC que nos ha permitido recorrer todas las fases del proceso **y conseguir nuestros tres objetivos:** la elaboración del corpus paralelo y la creación, puesta en marcha y evaluación de un motor de TA. En este sentido, la primera conclusión es que, efectivamente, MTUOC es proyecto integrador que permite a los usuarios —traductores, investigadores, empresas— con conocimientos prácticos del lenguaje y las técnicas de programación realizar todo el procedimiento sin demasiadas complicaciones. En el terreno de la investigación, el proyecto MTUOC se incluye entre las investigaciones más punteras para cubrir las demandas y necesidades de los estudios avanzados de traducción utilizando una metodología útil y que merece ser desarrollada.

En el capítulo anterior hemos detallado los resultados del proceso llevado a cabo, del cual podemos extraer las siguientes conclusiones:

1) Tras analizar los problemas en la ejecución de los entrenamientos que nos han impedido mejorar el rendimiento de nuestro motor, concluimos que el origen está en la insuficiente precisión de la alineación automática de nuestro corpus. Las causas de la baja calidad de la alineación del corpus son varias:

- En primer lugar, los textos descargados del sitio *Project Gutenberg* contienen extensos encabezados editoriales que se repiten al principio y final de cada obra; a su vez, los libros de la Biblioteca Nacional y Biblioteca Calibre contienen prefacios, notas o apéndices; también existen paralelismos incorrectos en algunas obras donde se mezclan el orden de los capítulos. Todo ello da lugar a alineaciones poco fiables que se pueden eliminar aumentando el índice de fiabilidad, pero con una pérdida importante de datos que deja el corpus vacío de contenido.
- En segundo lugar, la traducción literaria permite una buena dosis de libertad y creatividad por parte del traductor, pero esta característica impide que la correspondencia entre segmentos originales y traducidos sea muchas veces precisa con el consiguiente perjuicio que puede causar a la alineación del corpus y, por ende, al entrenamiento del sistema.

2) Partíamos de la hipótesis de que un motor de traducción entrenado específicamente con un corpus literario podría superar las prestaciones de un sistema genérico como Google Translate. Del análisis de los resultados de la evaluación podemos interpretar que nuestra hipótesis —sin llegar a corroborarse — se confirma como muy probable puesto que, a pesar de los problemas, nuestro motor es capaz de proporcionar traducciones comparables a los del sistema genérico. A su vez, se infiere que la puntuación otorgada a las traducciones depende de la calidad de la alineación en el texto de referencia.

En estas circunstancias, podemos afirmar que los resultados son relevantes porque permiten constatar que, si conseguimos mejorar la alineación del corpus central literario, mejoraremos el rendimiento de nuestro motor y con ello la calidad de los resultados de traducción.

5.2 Limitaciones del trabajo

- Nuestro único criterio a la hora de seleccionar las obras literarias que componen el corpus paralelo fue que cada una tuviera su correspondiente traducción. Se puede argumentar que no se adoptaron criterios de calidad de traducción adecuados a este tipo de texto y que confiamos excesivamente en la alineación

y selección automática, una técnica que ha revelado algunas carencias para conseguir la necesaria precisión en textos literarios.

- Por otra parte, se puede objetar que los resultados de la evaluación del sistema entrenado no son significantes puesto que provienen de datos erróneos y no definitivos. En todo caso, sirven para dar la medida de lo hecho hasta ahora y la comparación de sistemas es objetiva.
- Por último, somos conscientes de que la traducción literaria es un caso especial dentro de la TA que requiere una valoración por parte del lector. En este sentido, faltaría incluir una evaluación del usuario final sobre el grado de aceptabilidad y comprensibilidad que le otorgan a la TA en el contexto literario.

5.3 Futuras vías de investigación

Para una investigación posterior se pueden explorar, por ejemplo:

- Técnicas para limpiar el corpus automáticamente, por ejemplo, se podrían detectar los párrafos comunes entre libros y eliminarlos.
- Técnicas más avanzadas de alineación de documentos que filtren con más precisión los segmentos con una alineación poco fiable.
- Llevar a cabo una evaluación humana que valore el impacto en la lectura de un texto traducido con TA.

Agraïments

No puc tancar aquest TFM, i per extensió tot el màster, sense el meu reconeixement al professor Antoni Oliver per la seua disponibilitat i ajuda al llarg de tot aquest procés d'aprenentatge tan fascinant per a mi. També vull esmentar la meua tutora Silvia Rodríguez pels seus consells i suggeriments sempre adients i valuosos. I tampoc puc deixar d'agrair a Vicent que mai no s'ha queixat de la meua dedicació; Àlex i Ana de qui tan orgullosa estic, i la resta de la família i amics pel seu suport anímic.

A tots, mots gràcies!

6. Bibliografía

Bernardo, L. (2014). Tradução Literária homem VS máquina : um ensaio sobre a tradução online. *Entretextos, Londrina*, *14*(2), 191–204. Retrieved from https://cutt.ly/lfP43Tf

Bowker, L., & Pearson, J. (2002). Working with specialized language. A practical guide to using corpora. *Routledge, Londres*.

Briceño V., G. (2021). Lewis Carroll. Retrieved from https://www.euston96.com/lewis-carroll/

Briva-Iglesias, V. (2020). *Traducción automática inglés-catalán : tecnología de vanguardia , calidad y productividad*. *TFM*.

Costa Pellicer, O. (2021). *Estudio del enfoque web as corpus vs wb for corpus en la creción de un corpus ad hoc paralelo de resúmenes y abstracts sobre la intolerancia a la lactosa*.

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

Forcada, M. L., et al. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, *25*(2), 127–144. Retrieved from http://www.jstor.org/stable/41487458

Forcada, M. L., & Ñeco, R. P. (1997). Asynchronous translations with recurrent neural nets. *Proceedings of International Conference on Neural Networks (ICNN'97)*, *4*, 2535–2540 vol.4.

Giménez, J. (2008). Empirical Machine Translation and its Evaluation, 225. Retrieved from http://www.lsi.upc.edu/~jgimenez/PUBS/EMT.yahoo08.pdf

González Rey, M. I. (2014). Creación de un corpus literario paralelo como herramienta didáctica en fraseología bilingüe francés-español. *Fraseología y Paremiología*, 153–176. Retrieved from https://cvc.cervantes.es/lengua/biblioteca_fraseologica/n5_durante/gonzalez_rey_07.htm

Görög, A. (2014). Dynamic Quality Framework: quantifying and benchmarking quality. *Tradumàtica: Tecnologies de La Traducció*, (12), 443. https://doi.org/10.5565/rev/tradumatica.66

Guerberof, A., & Toral, A. (2021). The Impact if Pos-editing and Machine Translation on Creativity and Reading Experience. Retrieved from https://www.researchgate.net/publication/348563718

Guerberof , A., & Toral, A. (2022). *Creativity in translation: machine translation as a constraint for literary texts*.

Harris, B. (1988). Bi-text, a new concept in Translation theory. *Language Monthly*, *54*(Mars), 8–10.

House, J. (2012). Quality in translation studies. In *The Rutledge handbook of Translation and Technology*.

Hutchins, J. (2010). Machine translation: a concise history. *Journal of Translation Studies*, *13*(1–2), 29–70.

Jones, & Irvine. (2013). The (Un)faithful Machine Translator. *Proceeding of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 96–101.

Koby, et al. (2014). Defining Translation Quality. *Tradumàtica: Tecnologies de La Traducció*, *12*(413). Retrieved from https://doi.org/10.5565/rev/tradumatica.76

Koponen, M., Salmi, L., & Nikulin, M. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*. https://doi.org/10.1007/s10590-019-09228-7

Landers, C. E. (2001). *Literary translation a practical guide*. book, Clevedon ; Multilingual Matters. https://doi.org/10.21832/9781853595639

Leiva Rojo, J. (2018). Designing and compiling parallel aligned corpora: Pitfalls and (some) solutions on the example of a corpus of translated musem texts (English-Spanish). *Revista de Lingüística y Lenguas Aplicadas*, *13*(1), 59–73. article. https://doi.org/10.4995/rlyla.2018.7912

Matusov, E. (2019). The Challenges of Using Neural Machine Translation for Literature.

Proceedings of the Qualities of Literary Machine Translation, 10–19. Retrieved from http://matrix.statmt.org/matrix/systems_list/1914

Nádvorníková, O. (2017). Pièges méthodologiques des corpus parallèles et comment les éviter. *Corela: Cognition,Représentation, Langage*, (HS-21). https://doi.org/10.4000/corela.4810

Nunes Vieira, L. (2019). Post-editing of machine translation. In M. MinakoO'Hagan (Ed.), *The Rutledge handbook of Translation and Technology* (pp. 319–332).

Oliver, A. (2020). Corpus Paralelos. Retrieved from https://xwiki.recursos.uoc.edu/wiki/mat00001ca/view/1.4. Traducción automática y postedición/2. Traducción automática estadística con Moses/3. Corpus paralelos/

Oliver, A. (2021). Corpus UAB Alineación automática. Retrieved April 19, 2022, from https://youtu.be/zMT8TaKFjHc

Oliver, Antoni, & Alvarez-Vidal, S. (2021). Methods and Concepts for Researching Post-editings of Machine Translation. *7th IATIS Conference （The Cultural Ecology of Translation）*, (September). https://doi.org/10.13140/RG.2.2.15517.64485

Pym, A. (2013). Translation Skill-Sets in a Machine-Translation Age.

Rodriguez del Rosario, C. (2021). *Creación de motores de traducción automática (estadística y neuronal) inglés-español especializados en el campo de la aviación con la herramienta MTUOC*. Retrieved from http://hdl.handle.net/10609/133768

Saldanha, G., & O'Brien, S. (2014). *Research Methodologies in Translation Studies*. *Routledge*. Routledge. https://doi.org/10.4324/9781315760100

Santoyo, J. C. (1996). *El delito de traducir*. Publicaciones de la Universidad de León. Retrieved from https://fdocuments.es/document/el-delito-de-traducir.html?page=1

Serrano, R. (2020). Traducción automática y literatura: ¿enemigas íntimas? Retrieved from https://vasoscomunicantes.ace-traductores.org/2020/09/11

Simard, M. (2020). *Building and using parallel text for translation*. *The Routledge Handb. of Translation and Cognition*. https://doi.org/10.4324/9781315178127

Song Xiaoshu, C. D. (2003). Translation of Literary Style. Retrieved April 10, 2022, from http://www.accurapid.com/journal/23style.htm

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 2214–2218.

Toral, A., & Way, A. (2015). Translating Literary Text between Related Languages using SMT. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the 4th Workshop on Computational Linguistics for Literature, CLFL 2015*, 123–132. https://doi.org/10.3115/v1/w15-0714

Toral, A., & Way, A. (2018). What Level of Quality Can Neural Machine Translation Attain on Literary Text?, 263–287. https://doi.org/10.1007/978-3-319-91241-7_12

Udina, D. (2018). La traducción: una lectura exigente. *Vasos Comunicantes*, *48–49*. Retrieved from https://vasoscomunicantes.ace-traductores.org/2019/10/24/la-traduccion-una-lectura-exigente/

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2005). Parallel corpora for medium density languages. *International Conference Recent Advances in Natural Language Processing, RANLP*, *2005-Janua*(2003), 590–596. https://doi.org/10.1075/cilt.292.32var

Vaswani, & et al. (2017). Attention is All You Need. *31st Conference on Neural Information Processing System*. https://doi.org/10.1109/2943.974352

Youdale, R. (2020). *Using computers in the translation of literary style. Challenges and opportunities*. London and New York: Routledge.

Zanettin, F. (2012). *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. Taylor and Francis. Retrieved from https://doi.org/10.4324/9781315759661

Zanettin, F. (2014). Corpora in Translation. In J. House (Ed.), *Translation: A Multidisciplinary Approach* (pp. 178–199). London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137025487_10

Anexos

Anexo I: Listado de obras del corpus literario

A. A. Milne	The Red House Mystery
Agatha Christie	Poirot Investigates
Agatha Christie	The Man in the Brown Suit
Agatha Christie	The Murder on the Links
Agatha Christie	The Mysterious Affair at Styles
Agatha Christie	The Secret Adversary
Aleksandr Sergeevich Pushkin	Boris Godunov: a drama in verse
Aleksandr Sergeevich Pushkin	Eugene Oneguine
Aleksandr Sergeevich Pushkin	The Daughter of the Commandant
Alexandre Dumas	Man in the Iron Mask (an Essay)
Alexandre Dumas	The Countess of Charny;
Alexandre Dumas	The Forty-Five Guardsmen
Alexandre Dumas	The Queen's Necklace
Alexandre Dumas	The Three Musketeers
Alexandre Dumas	The Vicomte de Bragelonne
Alexandre Dumas	Twenty Years After
Algernon Blackwood	John Silence, Physician Extraordinary
Algernon Blackwood	The Empty House and Other Ghost Stories
Algernon Blackwood	The Wendigo
Algernon Blackwood	The Willows
Anatole France	Penguin Island
Anatole France	Thais
Anatole France	The Amethyst Ring
Anatole France	The Crime of Sylvestre Bonnard
Anatole France	The Procurator of Judea
Anatole France	The Revolt of the Angels
Anne Brontë	Agnes Grey
Anthony Hope	Rupert of Hentzau:
Anthony Hope	The Prisoner of Zenda
Aristophanes	Lysistrata
Aristophanes	The Birds
Aristotle	Politics: A Treatise on Government
Armando Palacio Valdés	Froth: A Novel
Armando Palacio Valdés	Maximina
Armando Palacio Valdés	The Joy of Captain Ribot
Arnold Bennett	Buried Alive: A Tale of These Days
Arnold Bennett	The Grand Babylon Hôtel
Arthur Conan Doyle	A Study in Scarlet
Arthur Conan Doyle	His Last Bow: An Epilogue of Sherlock Holmes
Arthur Conan Doyle	Round the Red Lamp
Arthur Conan Doyle	The Adventures of Sherlock Holmes
Arthur Conan Doyle	The Hound of the Baskervilles
Arthur Conan Doyle	The Memoirs of Sherlock Holmes
Arthur Conan Doyle	The Return of Sherlock Holmes
Arthur Conan Doyle	The Sign of the Four
Arthur Conan Doyle	The Tragedy of the Korosko
Arthur Conan Doyle	The Valley of Fear
Arthur Conan Doyle	The White Company
Arthur Machen	The Great God Pan
Arthur Machen	The Hill of Dreams
August Strindberg	The Red Room

August StrindbergThe Son of a Servant
Benito Pérez Galdós Dona Perfecta
Benito Pérez Galdós Leon Roch: A Romance, vol. 1 (of 2)
Benito Pérez Galdós Leon Roch: A Romance, vol. 2 (of 2)
Benito Pérez Galdós Marianela
Benito Pérez Galdós Saragossa: A Story of Spanish Valor
Benito Pérez Galdós The Novel on the Tram
Benito Pérez Galdós Trafalgar: A Tale
Benjamin FranklinThe Autobiography of Benjamin Franklin
Bernard Shaw Pygmalion
Bertrand Russell The Problems of Philosophy
Bram Stoker Dracula
Bram Stoker Dracula's Guest
Bram Stoker The Jewel of Seven Stars
Bram Stoker The Lair of the White Worm
Bruce Sterling The Hacker Crackdown
Carlo Collodi The Adventures of Pinocchio
Charles Baudelaire Poems in Prose
Charles Baudelaire The Flowers of Evil
Charles Dickens A Christmas Carol
Charles Dickens A Tale of Two Cities
Charles Dickens Bleak House
Charles Dickens David Copperfield
Charles Dickens Hard Times
Charles Dickens Little Dorrit
Charles Dickens Oliver Twist
Charles Dickens Our Mutual Friend
Charles Dickens Pictures from Italy
Charles Dickens The Haunted Man and the Ghost's Bargain
Charles Dickens The Mystery of Edwin Drood
Charlotte Brontë Jane Eyre: An Autobiography
Charlotte Brontë Villette
Christopher Morley The Haunted Bookshop
David Hume A Treatise of Human Nature
E. F. Benson Queen Lucia
Edgar Allan Poe Eureka: A Prose Poem
Edgar Allan Poe The Raven
Edgar Rice Burroughs A Princess of Mars
Edgar Rice Burroughs At the Earth's Core
Edgar Rice Burroughs Out of Time's Abyss
Edgar Rice Burroughs Pellucidar
Edgar Rice Burroughs Tarzan and the Ant Men
Edgar Rice Burroughs Tarzan and the Golden Lion
Edgar Rice Burroughs The Beasts of Tarzan
Edgar Rice Burroughs The Gods of Mars
Edgar Rice Burroughs The Land That Time Forgot
Edgar Rice Burroughs The Return of Tarzan
Edgar Rice Burroughs The Son of Tarzan
Edgar Wallace The Clue of the Twisted Candle
Edith Wharton Ethan Frome
Edith Wharton Madame de Treymes
Edith Wharton Sanctuary
Edith Wharton The Age of Innocence
Edith Wharton The Touchstone
F. Scott Fitzgerald The Beautiful and Damned
F. Scott Fitzgerald This Side of Paradise
Frances Hodgson Burnett A Little Princess
Frances Hodgson Burnett The Secret Garden

Franz Kafka Metamorphosis
Franz Kafka The Trial
Fritz Leiber The Big Time
Frédéric Bastiat The Law
G. K. Chesterton A Short History of England
G. K. Chesterton Manalive
G. K. Chesterton Orthodoxy
G. K. Chesterton The Club of Queer Trades
G. K. Chesterton The Flying Inn
G. K. Chesterton The Innocence of Father Brown
G. K. Chesterton The Man Who Was Thursday: A Nightmare
G. K. Chesterton The Wisdom of Father Brown
George Eliot Brother Jacob
George Eliot Silas Marner
George Eliot The Lifted Veil
George Eliot The Mill on the Floss
Giovanni Boccaccio The Decameron of Giovanni Boccaccio
Gustave Flaubert Madame Bovary
Gustave Flaubert Three short works
Gustavo A. Bécquer Romantic legends of Spain
Guy de Maupassant Bel Ami; Or, The History of a Scoundre
Guy de Maupassant Mont Oriol; or, A Romance of Auvergne
H. P. Lovecraft The Dunwich Horror
H. Rider Haggard Allan Quatermain
H. Rider Haggard Ayesha, the Return of She
H. Rider Haggard Child of Storm
H. Rider Haggard King Solomon's Mines
H. Rider Haggard She
H. Rider Haggard The Ghost Kings
H. Rider Haggard The People of the Mist
H. Rider Haggard Wisdom's Daughter:
Hans Jakob Christoph von Grimmelshausen Simplicissimus
Harry Harrison Deathworld
Henri Barbusse The Inferno
Henri Bergson Creative Evolution
Henri Bergson Laughter: An Essay on the Meaning of the Comic
Henrik Ibsen A Doll's House
Henrik Ibsen An Enemy of the People
Henry David Thoreau Walking
Henry James Daisy Miller: A Study
Henry James In the Cage
Henry James Madame De Mauves
Henry James The Ambassadors
Henry James The American
Henry James The Diary of a Man of Fifty
Henry James The Europeans
Henry James The Lesson of the Master
Henry James The Turn of the Screw
Henry James Washington Square
Henry James What Maisie Knew
Herman Melville Bartleby, the Scrivener: A Story of Wall-Street
Herman Melville Moby-Dick; or, The Whale
Herman Melville Moby Dick
Herman Melville Pierre; or The Ambiguities
Herman Melville Typee
Hermann Hesse Siddhartha
Honoré de Balzac Albert Savarus
Honoré de Balzac Cousin Pons

Honoré de Balzac Eugenie Grandet
Honoré de Balzac Lost Illusions
Honoré de Balzac The Lesser Bourgeoisie
Honoré de Balzac The Lily of the Valley
Horacio Quiroga South American Jungle Tales
Horacio Quiroga South American Jungle Tales
Howard Pyle The Merry Adventures of Robin Hood
Hugh Lofting Doctor Dolittle's Post Office
Hugh Lofting The Story of Doctor Dolittle
Jack London Adventure
Jack London Martin Eden
Jack London The Call of the Wild
Jack London The Cruise of the Snark
Jack London The Iron Heel
Jack London The People of the Abyss
Jack London The Son of the Wolf
Jack London The Valley of the Moon
Jack London White Fang
James Joyce Dubliners
James Joyce Ulysses
Jane Austen Emma
Jane Austen Lady Susan
Jane Austen Love and Freindship [sic]
Jane Austen Pride and Prejudice
Jane Austen Sense and Sensibility
Jean Racine Phaedra
Jean-Jacques Rousseau Émile;
Johann Wolfgang von Goethe Hermann and Dorothea
Johann Wolfgang von Goethe The Sorrows of Young Werther
Johanna Spyri Heidi
John Buchan The Thirty-Nine Steps
John Dos Passos Rosinante to the Road Again
John Dos Passos Three Soldiers
John Galsworthy The Forsyte Saga - Complete
John Keats Lamia
John Reed Ten Days That Shook the World
John Stuart Mill On Liberty
Jonathan Swift A Modest Proposal
Jonathan Swift Gulliver's Travels
Joseph Conrad Heart of Darkness
Joseph Conrad Lord Jim
Joseph Conrad Nostromo: A Tale of the Seaboard
Joseph Conrad Tales of Unrest
Joseph Conrad The Mirror of the Sea
Joseph Conrad The Nigger Of The "Narcissus": A Tale Of The Forecastle
Joseph Conrad The Secret Sharer
Joseph Conrad The Shadow Line: A Confession
Joseph Conrad Victory: An Island Tale
José Rizal The Philippines a Century Hence
José Rizal The Social Cancer: José Zorrilla Don Juan Tenorio
Juan Valera Pepita Ximenez
Juan Valera Pepita Ximenez
Keith Laumer A Trace of Memory
L. M. Montgomery Anne of Avonlea
L. M. Montgomery Anne of Green Gables
L. M. Montgomery Anne's House of Dreams
L. M. Montgomery Emily of New Moon

L. M. Montgomery Rilla of Ingleside
Leo Tolstoy Anna Karenina
Leo Tolstoy Katia
Leo Tolstoy Resurrection
Leo Tolstoy War and Peace
Leo Tolstoy What Shall We Do?
Lewis Carroll Alice's Adventures in Wonderland
Lewis Carroll Alice's Adventures in Wonderland
Lewis Carroll The Hunting of the Snark: An Agony, in Eight Fits
Lewis Carroll Through the Looking-Glass
Lord Dunsany A Dreamer's Tales
Louisa May Alcott Little Men: Life at Plumfield With Jo's Boys
Louisa May Alcott Little Women; Or, Meg, Jo, Beth, and Amy
Louisa May Alcott Under the Lilacs
Marie Lebert From the Print Media to the Internet
Marion Zimmer Bradley The Door Through Space
Mark Twain A Connecticut Yankee in King Arthur's Court
Mark Twain Adventures of Huckleberry Finn
Mark Twain The Adventures of Tom Sawyer, Complete
Mark Twain Tom Sawyer, Detective
Maurice Leblanc 813
Maurice Leblanc Arsène Lupin versus Herlock Sholmes
Maurice Leblanc The Crystal Stopper
Maurice Leblanc The Hollow Needle; Further adventures of Arsene Lupin
Maurice Maeterlinck The Life of the Bee
Miguel de Cervantes Saavedra Don Quixote
Miguel de Cervantes Saavedra The Exemplary Novels of Cervantes
Miguel de Unamuno Tragic Sense Of Life
Molière Tartuffe; Or, The Hypocrite
Molière The Imaginary Invalid
Molière The Learned Women
Molière The Miser
Nathaniel Hawthorne Tanglewood Tales
Nathaniel Hawthorne The House of the Seven Gables
Nathaniel Hawthorne The Scarlet Letter
Nathaniel Hawthorne Twice Told Tales
Oscar Wilde A House of Pomegranates
Oscar Wilde A Woman of No Importance
Oscar Wilde De Profundis
Oscar Wilde Salomé: A Tragedy in One Act
Oscar Wilde The Happy Prince, and Other Tales
Oscar Wilde The Importance of Being Earnest:
P. G. Wodehouse The Man with Two Left Feet, and Other Stories
Pedro Calderón de la Barca Life Is a Dream
Pío Baroja Cæsar or Nothing
Pío Baroja The Quest
Pío Baroja Weeds
Rafael Sabatini Captain Blood
Rafael Sabatini Scaramouche
Ralph Waldo Emerson The Conduct of Life
Ray Cummings Brigands of the Moon
Robert Hugh Benson Lord of the World
Robert Louis Stevenson Catriona
Robert Louis Stevenson Kidnapped
Robert Louis Stevenson The Black Arrow: A Tale of the Two Roses

Robert Louis Stevenson The Master of Ballantrae: A Winter's Tale
Robert Louis Stevenson The Strange Case of Dr. Jekyll and Mr. Hyde
Robert Louis Stevenson Treasure Island
Robert Sheckley Meeting of the Minds
Robert Sheckley The Hour of Battle
Robert W. Chambers The King in Yellow
Rudyard Kipling "Captains Courageous": A Story of the Grand Banks
Rudyard Kipling Kim
Rudyard Kipling Puck of Pook's Hill
Rudyard Kipling Stalky & Co.
Rudyard Kipling The Bridge-Builders
Rudyard Kipling The Man Who Would Be King
Rudyard Kipling The Second Jungle Book
Saki Beasts and Super-Beasts
Saki The Chronicles of Clovis
Samuel Butler Erewhon; Or, Over the Range
Samuel Johnson Preface to Shakespeare
Selma Lagerlöf Jerusalem
Selma Lagerlöf The Story of Gösta Berling
Sherwood Anderson Poor White: A Novel
Sherwood Anderson Winesburg, Ohio
Sinclair Lewis Babbitt
Sinclair Lewis Free Air
Sinclair Lewis Main Street
Sophocles Philoktetes
Sophocles Plays: Oedipus the King; Oedipus at Colonus; Antigone
Stefan Zweig Jeremiah: A Drama in Nine Scenes
Stefan Zweig The Burning Secret
Stendhal The Chartreuse of Parma
Stephen Crane Maggie: A Girl of the Streets
Stephen Crane The Red Badge of Courage:
Theodor Mommsen The History of Rome, Book III
Theodor Mommsen The History of Rome, Book IV
Theodor Mommsen The History of Rome, Book V
Thomas De Quincey Confessions of an English Opium-Eater
Thomas De Quincey De Quincey's Revolt of the Tartars
Thomas Hardy A Pair of Blue Eyes
Thomas Hardy The Mayor of Casterbridge
Thomas Hardy The Return of the Native
Thomas Hobbes Leviathan
Théophile Gautier Wanderings in Spain
Tomás de Iriarte Literary Fables of Yriarte
Vicente Blasco Ibáñez Mare Nostrum (Our Sea): A Novel
Vicente Blasco Ibáñez The Blood of the Arena
Vicente Blasco Ibáñez The Cabin [La barraca]
Vicente Blasco Ibáñez The Dead Command
Vicente Blasco Ibáñez The Enemies of Women (Los enemigos de la mujer)
Vicente Blasco Ibáñez The Four Horsemen of the Apocalypse
Vicente Blasco Ibáñez The Shadow of the Cathedral
Vicente Blasco Ibáñez The Torrent (Entre Naranjos)
Vicente Blasco Ibáñez Woman Triumphant (La Maja Desnuda)
Voltaire Candide
Voltaire Candide
Voltaire Micromegas
Voltaire Voltaire's Romances, Complete in One Volume
Voltaire Zadig; Or, The Book of Fate

W. B. Yeats The Celtic Twilight
W. Somerset Maugham The Painted Veil
Walt Whitman Leaves of Grass
Walter Lippmann Public Opinion
Ward Moore Greener Than You Think
Washington IrvingThe Alhambra
Wassily KandinskyConcerning the Spiritual in Art
Wilkie Collins Armadale
Wilkie Collins Basil
Wilkie Collins Jezebel's Daughter
Wilkie Collins Man and Wife
Wilkie Collins My Lady's Money
Wilkie Collins No Name
Wilkie Collins Poor Miss Finch
Wilkie Collins The Fallen Leaves
Wilkie Collins The Frozen Deep
Wilkie Collins The Guilty River
Wilkie Collins The Legacy of Cain
Wilkie Collins The Moonstone
Wilkie Collins The Queen of Hearts
Wilkie Collins The Two Destinies
Wilkie Collins The Woman in White
William Hope Hodgson Carnacki, the Ghost Finder
William Hope Hodgson The Ghost Pirates
William Hope Hodgson The House on the Borderland
William Morris The Wood Beyond the World
William Shakespeare A Midsummer Night's Dream
William Shakespeare The Merchant of Venice
William Shakespeare The Merry Wives of Windsor
William Shakespeare The Tempest
William Shakespeare The Tragedy of Antony and Cleopatra
William Shakespeare The Tragedy of Hamlet, Prince of Denmark
William Shakespeare The Tragedy of Julius Caesar
William Shakespeare The Tragedy of King Lear
William Shakespeare The Tragedy of King Richard III
William Shakespeare The Tragedy of Macbeth
William Shakespeare The Tragedy of Othello, Moor of Venice
William Shakespeare The Tragedy of Romeo and Juliet
William Tenn Of All Possible Worlds
Zane Grey Riders of the Purple Sage
Zane Grey The Border Legion
Zane Grey The Call of the Canyon
Zane Grey The Desert of Wheat
Zane Grey The Last Trail
Zane Grey The Man of the Forest
Zane Grey The Mysterious Rider
Zane Grey The Rainbow Trail
Zane Grey The Spirit of the Border
Zane Grey To the Last Man

Anexo II: Corpus de validación

a) Fragmento del corpus

s> _sorrows _and _hopes _. </s> <s> _La _noche _de _aquel _día _trajo _el _invierno _. </s>
<s> _Another _day _brought _dull _- _gray _sc ud ding _cloud_, _and _gust s _of _wind _and </s><s> _A _la _mañana _siguiente _Wade _fue _a _ver _a _Wilson _Moore _march ando _sobre _una _alfombra _de _nieve _de _dos _pal mos _de _espesor _. </s>
<s> _Next _morning _, _when _Wade _plo d ded _up _to _Moore _' s _cabin _, _it _was _through _two </s> <s> _— _Buenos _días _, _Wilson _— _dijo _Wade _sacud iéndose _la _nieve _de _las _botas _— _. </s>
<s> _When _Wade _pushed _open _the _door _of _the _cabin _and _entered _he _awakened _the </s> <s> _— _Buenos _días _, _Wade _— _contestó _Moore _— _. </s>
<s> _" _Well _, _I _was _worried _about _that _, _" _said _the _hunter _. </s> <s> _Dos _pal mos _de _nieve _en _el _suelo _hay _esta _mañana _. </s>
<s> _" _We _' ve _got _to _arrange </s> <s> _Por _fortuna _, _aquí _cerca _tengo _preparada _una _gran _cantidad _de _leña _. </s>
<s> _" _Won _' t _Old _Bill _make _a _kick _? _" </s><s> _— _Si _se _enf ada _, _que _se _enf ade _. </s>
<s> _Lem _got _at _Kre mm lin _' _the _other _day _. </s> <s> _Aquí _tienes _algunas _cartas _que _Manuel _trajo _de _Kre mm ling _el _otro _día _. </s>
<s> _Moore _scanned _the _addresses _on _the _several _envelop es _and _sighed _. </s> <s> _Moore _echó _una _ojeada _a _los _varios _sobres _y _lanzó _un _suspiro _. </s>
<s> _I _hate _to _read _them _. _" </s> <s> _— _¡ _Cart as _de _casa _! </s>
<s> _and _my _sister _. </s> <s> _— _Pues _he _hecho _, _sí _, _he _hecho _una _porción _de _cosas _. </s>
<s> _couldn _' t - _- _I _couldn _' t _ever _ride _a _horse _again - _- _if _I _did _go _. _" </s> <s> _Quiere _volver me _a _ver _, _me _perdona _. </s>
<s> _" _I _never _said _so _. </s> <s> _¿ _No _es _tremendo _? </s>
<s> _somewhat _the _burden _of _Wade _' s _worry _. </s> <s> _Much acho _, _me _acabas _de _dar _una _buena _nueva _. </s>
<s> _Wade _, _when _the _dreaded _time _could _be _put _off _no _longer _. </s> <s> _Adelante _, _Wade _, _y _si _ves _que _mi _pierna _no _puede _salvarse _, _alc án z ame _el _revólver _. </s>

b) Tabla con los primeros segmentos en el que podemos apreciar los errores de alineación

<s> _sorrows _and _hopes _. </s>	<s> _La _noche _de _aquel _día _trajo _el _invierno _. </s>
<s> _Another _day _brought _dull _- _gray _sc ud ding _clouds _, _and _gust s _of _wind _and </s>	<s> _A _la _mañana _siguiente _Wade _fue _a _ver _a _Wilson _Moore _marchando _sobre _una _alfombra _de _nieve _de _dos _palmos _de _espesor _. </s>
<s> _Next _morning _, _when _Wade _plo d ded _up _to _Moore _' s _cabin _, _it _was _through _two </s>	<s> _— _Buenos _días _, _Wilson _— _dijo _Wade _sacudiéndose _la _nieve _de _las _botas _— _. </s>
<s> _When _Wade _pushed _open _the _door _of _the _cabin _and _entered _he _awakened _the </s>	<s> _— _Buenos _días _, _Wade _— _contestó _Moore _— _. </s>
_Moore _scanned _the _addresses _on _the _several _envelop es _and _sighed _. </s>	<s> _Moore _echó _una _ojeada _a _los _varios _sobres _y _lanzó _un _suspiro _. </s>
<s> _I _hate _to _read _them _. _" </s>	<s> _— _¡ _Cartas _de _casa _! </s>
<s> _and _my _sister _. </s>	<s> _— _Pues _he _hecho _, _sí _, _he _hecho _una _porción _de _cosas _. </s>
<s> _couldn _' t - _- _I _couldn _' t _ever _ride _a _horse _again - _- _if _I _did _go _. _" </s>	_Quiere _volver me _a _ver _, _me _perdona _. </s>
<s> _" _I _never _said _so _. </s>	<s> _¿ _No _es _tremendo _? </s>

c) Archivo *valid.log* donde quedan registradas las incidencias del entrenamiento

```
[2022-05-13 09:48:36] [valid] Ep. 1 : Up. 5000 : cross-entropy : 121.404 : new best
[2022-05-13 09:48:43] [valid] First sentence's tokens as scored:
[2022-05-13 09:48:43] [valid] DefaultVocab keeps original segments for scoring
[2022-05-13 09:48:43] [valid]   Hyp: <s> ▁En ▁cuanto ▁a ▁la ▁puerta ▁, ▁y ▁a ▁la
▁puerta ▁de ▁la ▁señora ▁.
[2022-05-13 09:48:43] [valid]   Ref: <s> ▁Durante ▁unos ▁minutos ▁quedó ▁fuera ▁de
▁las ▁miradas ▁de ▁Benj amín ▁Wade ▁, ▁para ▁volver ▁a ▁aparecer ▁luego
▁montado ▁en ▁un ▁caballo ▁blanco ▁con ▁el ▁cual ▁se ▁dirigió ▁por ▁los
▁prados ▁a ▁la ▁ladera ▁de ▁la ▁montaña ▁, ▁desde ▁donde ▁marchó ▁hacia ▁un
▁o qu ed al ▁bastante ▁cercano ▁a ▁la ▁cabaña ▁de ▁Wilson ▁Moore ▁.
[2022-05-13 09:56:06] [valid] Ep. 1 : Up. 5000 : bleu-detok : 1.38611 : new best
[2022-05-13 12:17:40] [valid] Ep. 1 : Up. 10000 : cross-entropy : 103.04 : new best
[2022-05-13 12:23:40] [valid] Ep. 1 : Up. 10000 : bleu-detok : 4.20415 : new best
[2022-05-13 14:43:19] [valid] Ep. 1 : Up. 15000 : cross-entropy : 92.8762 : new best
[2022-05-13 14:48:56] [valid] Ep. 1 : Up. 15000 : bleu-detok : 6.66172 : new best
[2022-05-13 17:13:04] [valid] Ep. 1 : Up. 20000 : cross-entropy : 87.701 : new best
[2022-05-13 17:17:33] [valid] Ep. 1 : Up. 20000 : bleu-detok : 7.46365 : new best
[2022-05-13 19:42:40] [valid] Ep. 1 : Up. 25000 : cross-entropy : 84.3006 : new best
[2022-05-13 19:47:21] [valid] Ep. 1 : Up. 25000 : bleu-detok : 8.07549 : new best
[2022-05-13 22:13:03] [valid] Ep. 2 : Up. 30000 : cross-entropy : 82.0933 : new best
[2022-05-13 22:17:51] [valid] Ep. 2 : Up. 30000 : bleu-detok : 8.8294 : new best
[2022-05-14 00:46:33] [valid] Ep. 2 : Up. 35000 : cross-entropy : 80.4652 : new best
[2022-05-14 00:51:28] [valid] Ep. 2 : Up. 35000 : bleu-detok : 8.9188 : new best
[2022-05-14 03:23:03] [valid] Ep. 2 : Up. 40000 : cross-entropy : 79.4145 : new best
[2022-05-14 03:28:02] [valid] Ep. 2 : Up. 40000 : bleu-detok : 8.99877 : new best
[2022-05-14 06:01:16] [valid] Ep. 2 : Up. 45000 : cross-entropy : 78.5091 : new best
[2022-05-14 06:06:29] [valid] Ep. 2 : Up. 45000 : bleu-detok : 9.23201 : new best
```

Anexo III: Set de evaluación

a) Fragmento del texto original (The curious incident…)

“Why were you holding the dog?” he asked again.
“I like dogs,” I said.
“Did you kill the dog?” he asked.
I said, “I did not kill the dog.”
“Is this your fork?” he asked.
I said, “No.”
“You seem very upset about this,” he said.
He was asking too many questions and he was asking them too quickly.
They were stacking up in my head like loaves in the factory where Uncle Terry works.
The factory is a bakery and he operates the slicing machines.
And sometimes a slicer is not working fast enough but the bread keeps coming and there is a blockage.
I sometimes think of my mind as a machine, but not always as a bread-slicing machine.
It makes it easier to explain to other people what is going on inside it.
The policeman said, “I am going to ask you once again.

I find people confusing.
This is for two main reasons.
The first main reason is that people do a lot of talking without using any words.
Siobhan says that if you raise one eyebrow it can mean lots of different things.
It can mean “I want to do sex with you” and it can also mean “I think that what you just said was very stupid.”
The second main reason is that people often talk using metaphors.
These are examples of metaphors
I laughed my socks off.
He was the apple of her eye.
They had a skeleton in the cupboard.
We had a real pig of a day.

The word metaphor means carrying something from one place to another, and it comes from the Greek words meta (which means from one place to another) and ferein (which means to carry), and it is when you describe something by using a word for something that it isn't.
This means that the word metaphor is a metaphor.
I think it should be called a lie because a pig is not like a day and people do not have skeletons in their cupboards.
And when I try and make a picture of the phrase in my head it just confuses me because imagining an apple in someone's eye doesn't have anything to do with liking someone a lot and it makes you forget what the person was talking about.
My name is a metaphor.
It means carrying Christ and it comes from the Greek words χρίστοζ (which means Jesus
This makes you wonder what he was called before he carried Christ across the river.
But he wasn't called anything because this is an apocryphal story, which means that it is a lie, too.
Mother used to say that it meant Christopher was a nice name because it was a story about being kind and helpful, but I do not want my name to mean a story about being kind and helpful.
I want my name to mean me.

b) Fragmento de la traducción de referencia

— ¿Por qué tenías al perro en brazos? —preguntó otra vez.
— Me gustan los perros —dije.
— ¿Has matado al perro? —preguntó.
— Yo no he matado al perro —dije.
— ¿La horca es tuya? —preguntó.
— No —dije.
— Parece que esto te ha alterado mucho —dijo.
Me estaba haciendo demasiadas preguntas y me las estaba haciendo demasiado rápido.
Se me amontonaban como los panes en la fábrica donde trabaja el tío Terry.
La fábrica es una panificadora y él maneja la máquina de rebanar.
A veces la máquina no va lo bastante rápido pero el pan sigue llegando hasta causar un bloqueo.
A veces me imagino mi mente como si fuera una máquina, aunque no siempre como una rebanadora de pan.
Hace que me sea más fácil explicarles a los demás lo que pasa en mi interior.
El policía dijo: "Te lo volveré a preguntar".

La gente me provoca confusión.
Eso me pasa por dos razones principales.
La primera razón principal es que la gente habla mucho sin utilizar ninguna palabra.
Siobhan dice que si uno arquea una ceja puede querer decir montones de cosas distintas.
Puede significar «quiero tener relaciones sexuales contigo» y también puede querer decir «creo que lo que acabas de decir es
Siobhan también dice que si cierras la boca y expeles aire con fuerza por la nariz puede significar que estás relajado, o que estás aburrido o que estás enfadado, y todo depende de cuánto aire te salga por la nariz y con qué rapidez y de qué forma tenga tu boca cuando lo hagas y de cómo estés sentado y de lo que hayas dicho justo antes y de cientos de otras cosas que son demasiado complicadas para entenderlas en sólo unos segundos.
La segunda razón principal es que la gente con frecuencia utiliza metáforas.
He aquí ejemplos de metáforas
Me he muerto de risa
Era la niña de sus ojos
Tenían un cadáver en el armario
Pasamos un día de mil demonios

La palabra metáfora significa llevar algo de un sitio a otro, y viene de las palabras griegas μετα (que significa de un sitio a otro) y φερειν (que significa llevar), y es cuando uno describe algo usando una palabra que no es literalmente lo que describe.
Es decir, que la palabra metáfora es una metáfora.
Yo creo que debería llamarse mentira porque no hay días de mil demonios y la gente no tiene cadáveres en los armarios.
Cuando trato de formarme una imagen en mi cabeza de una de estas frases me siento perdido porque una niña en los ojos de alguien no tiene nada que ver con que algo le guste mucho y te olvidas de lo que la persona decía.
Mi nombre es una metáfora.
Significa «que lleva a Cristo» y viene de las palabras griegas χριστος ; (que significa Jesucristo) y φερειν , y fue el nombre que le pusieron a san Cristóbal porque cruzó un río llevando a Jesucristo.

Eso te hace pensar en cómo se llamaría Cristóbal antes de cruzar el río con Jesucristo a cuestas.
Pero no se llamaba de ninguna manera porque ésa es una historia apócrifa, lo cual significa que es, también, una mentira.
Madre solía decir que Christopher era un nombre bonito, porque es una historia sobre ser amable y servicial, pero yo no quiero que mi nombre se refiera a una historia sobre ser amable y servicial.
Yo quiero que mi nombre se refiera a mí.

c) Fragmento de la traducción con MTUOC

— ¿Por qué llevabas al perro? — preguntó de nuevo.
— Me gustan los perros — dije.
— ¿Mató usted al perro? — preguntó.
— Yo no maté al perro.
— ¿Es éste el tenedor? — preguntó.
— No — dije —.
— Parece usted muy disgustado — dijo —.
Le hacía demasiadas preguntas y se las preguntaba rápidamente.
Ellos estaban apilando en mi cabeza como panes en la fábrica donde el tío Terry trabaja.
La fábrica es una panadería y opera las máquinas de corte.
Y a veces un rebanador no funciona lo suficientemente rápido, pero el pan sigue llegando y hay un bloqueo.
A veces pienso en mi mente como una máquina, pero no siempre como una máquina de cortar pan.
Es más fácil explicar a otras personas lo que sucede dentro de ella.
— Te voy a preguntar una vez más.

Encuentro a personas que confunden.
Se trata de dos razones principales.
La primera razón principal es que la gente habla mucho sin usar palabras.
Siobhan dice que si levantas una ceja puede significar muchas cosas diferentes.
Puede significar “quiero hacer sexo con usted” y también puede significar “creo que lo que acaba de decir fue muy estúpido”.
La segunda razón principal es que las personas a menudo hablan con metáforas.
Estos son ejemplos de metáforas
Me reí mis calcetines.
Era la manzana de sus ojos.
Tenía un esqueleto en el armario.
Teníamos un verdadero cerdo de un día.

La palabra metáfora significa llevar algo de un lugar a otro, y viene de las palabras griegas meta (que significa de un lugar a otro) y ferein (que significa llevar), y es cuando describes algo usando una palabra para algo que no lo es.
Esto significa que la palabra metáfora es una metáfora.
Creo que debe ser llamado mentira porque un cerdo no es como un día y la gente no tiene esqueletos en sus armarios.
Y cuando trato de hacer una imagen de la frase en mi cabeza sólo me confunde porque imaginar una manzana en los ojos de alguien no tiene nada que ver con el gusto de alguien mucho y te hace olvidar lo que la persona estaba hablando.
Mi nombre es una metáfora.

Esto significa llevar a Cristo y proviene de las palabras griegas
Esto hace que te pregunte cómo fue llamado antes de llevar a Cristo a través del río.
Pero no fue llamado nada porque se trata de una historia apócrifa, lo que significa que es una mentira también.
Mamá solía decir que quería decir que Christopher era un buen nombre porque era una historia sobre ser amable y útil, pero no quiero que mi nombre se refiera a una historia sobre ser amable y útil.
Quiero que mi nombre se refiera a mí.

---

d) Fragmento de la traducción de Google Translate

"¿Por qué estabas sosteniendo al perro?" preguntó de nuevo.
"Me gustan los perros", dije.
"¿Mataste al perro?" preguntó.
Le dije: "No maté al perro".
"¿Es esta tu bifurcación?" preguntó.
Dije que no."
"Pareces muy molesto por esto", dijo.
Estaba haciendo demasiadas preguntas y las estaba haciendo demasiado rápido.
Se apilaban en mi cabeza como panes en la fábrica donde trabaja el tío Terry.
La fábrica es una panadería y opera las máquinas de corte.
Y a veces una cortadora no funciona lo suficientemente rápido, pero el pan sigue llegando y hay un bloqueo.
A veces pienso en mi mente como una máquina, pero no siempre como una máquina de sencillo de pan.
Hace que sea más fácil explicar a otras personas lo que está sucediendo dentro de él.
El policía dijo: “Te voy a preguntar una vez más.

Encuentro gente confusa.
Esto es por dos razones principales.
La primera razón principal es que las personas hablan mucho sin usar ninguna palabra.
Siobhan dice que si levantas una ceja, puede significar muchas cosas diferentes.
Puede significar "Quiero hacer sexo contigo" y también puede significar "Creo que lo que acabas de decir fue muy estúpido".
La segunda razón principal es que las personas a menudo hablan usando metáforas.
Estos son ejemplos de metáforas
Me reí mis calcetines.
Él era la manzana de su ojo.
Tenían un esqueleto en el armario.
Tuvimos un verdadero cerdo de día.

La palabra metáfora significa llevar algo de un lugar a otro, y proviene de las palabras griegas meta (que significa de un lugar a otro) y ferein (que significa llevar), y es cuando describe algo usando una palabra para algo que no es.
Esto significa que la palabra metáfora es una metáfora.
Creo que debería llamarse mentira porque un cerdo no es como un día y la gente no tiene esqueletos en sus armarios.

Y cuando trato de hacer una foto de la frase en mi cabeza, simplemente me confunde porque imaginar una manzana en el ojo de alguien no tiene nada que ver con gustarle mucho a alguien y te hace olvidar de qué estaba hablando la persona.
Mi nombre es una metáfora.
Significa llevar a Cristo y proviene de las palabras griegas χρίστοζ (lo que significa Jesús
Esto te hace preguntarte cómo se llamaba antes de llevar a Cristo a través del río.
Pero no se llamaba nada porque esta es una historia apócrifa, lo que significa que también es una mentira.
Madre solía decir que significaba que Christopher era un buen nombre porque era una historia sobre ser amable y servicial, pero no quiero que mi nombre signifique una historia sobre ser amable y servicial.
Quiero que mi nombre me referirá.

Anexo IV: Set de evaluación

a) Fragmento del texto original (Cap.6 de The Hobbit)

Chapter 6: Out of the Frying Pan Into the Fire
Bilbo had escaped the goblins, but he did not know where he was.
He had lost hood, cloak, food, pony, his buttons and his friends.
He wandered on and on, till the sun began to sink westwards-behind the mountains.
Their shadows fell across Bilbo's path, and he looked back.
Then he looked forward and could see before him only ridges and slopes falling towards lowlands and plains glimpsed occasionally between the trees.
"Good heavens!" he exclaimed.
"I seem to have got right to the other side of the Misty Mountains, right to the edge of the Land Beyond!
Where and O where can Gandalf and the dwarves have got to?
I only hope to goodness they are not still back there in the power of the goblins!"
He still wandered on, out of the little high valley, over its edge, and down the slopes beyond; but all the while a very uncomfortable thought was growing inside him.
He wondered whether he ought not, now he had the magic ring, to go back into the horrible, horrible, tunnels and look for his friends.
He had just made up his mind that it was his duty, that he must turn back-and very miserable he felt about it-when he heard voices.
He stopped and listened.
It did not sound like goblins; so he crept forward carefully.
He was on a stony path winding downwards with a rocky wall. on the left hand; on the other side the ground sloped away and there were dells below the level of the path overhung with bushes and low trees.
In one of these dells under the bushes people were talking.
He crept still nearer, and suddenly he saw peering between two big boulders a head with a red hood on: it was Balin doing look-out.
He could have clapped and shouted for joy, but he did not.
He had still got the ring on, for fear of meeting something unexpected and unpleasant, and he saw that Balin was looking straight at him without noticing him.
"I will give them all a surprise," he thought, as he crawled into the bushes at the edge of the dell.
Gandalf was arguing with the dwarves.
They were discussing all that had happened to them in the tunnels, and wondering and debating what they were to do now.
The dwarves were grumbling, and Gandalf was saying that they could not possibly go on with their journey leaving Mr. Baggins in the hands of the goblins, without trying to find out if he was alive or dead, and without trying to rescue him.
"After all he is my friend," said the wizard, "and not a bad little chap.
I feel responsible for him.
I wish to goodness you had not lost him."
The dwarves wanted to know why he had ever been brought at all, why he could not stick to his friends and come along with them, and why the wizard had not chosen someone with more sense.
"He has been more trouble than use so far,"said one.
"If we have got to go back now into those abominable tunnels to look for him, then drat him, I say."
Gandalf answered angrily: "I brought him, and I don't bring things that are of no use.

b) Fragmento del texto de referencia

6. De la sartén al fuego

Bilbo había escapado de los trasgos, pero no sabía dónde estaba.

Había perdido el capuchón, la capa, la comida, el poney, sus botones y sus amigos.

Siguió adelante, hasta que el sol empezó a hundirse en el poniente, detrás de las montañas.

Las sombras cruzaban el sendero, y Bilbo miró hacia atrás,

luego miró hacia adelante, y no pudo ver más que crestas y vertientes que descendían hacia las tierras bajas, y llanuras que asomaban de vez en cuando entre los árboles.

¡Cielos! exclamó.

¡Parece que estoy justo al otro lado de las Montañas Nubladas, al borde de las Tierras de Más Allá!

¿Dónde y adónde habrán tenido que ir los enanos y Gandalf?

¡Sólo espero que por ventura no estén todavía allá atrás en poder de los trasgos!

Continuó caminando, fuera del pequeño y elevado valle, por el borde, y bajando luego las pendientes; mas en todo este tiempo un pensamiento muy incómodo iba creciendo dentro de él.

Se preguntaba si no estaba obligado, ahora que tenía el anillo mágico, a regresar a los horribles, horribles túneles y buscar a sus amigos.

Acababa de decidir que no podía escapar a ese deber, que tenía que volver atrás y esto hacía que se sintiera muy desdichado cuando oyó voces.

Se detuvo y escuchó.

No parecían trasgos; de modo que se arrastró con mucho cuidado hacia adelante.

Estaba en un sendero pedregoso que serpenteaba hacia abajo, con una pared rocosa a la izquierda; al otro lado el terreno descendía en pendiente, y bajo el nivel del sendero había unas cañadas donde crecían matorrales y arbustos.

En una de estas cañadas, bajo los arbustos, había gente hablando

Se arrastró todavía más cerca, y de súbito vio, asomado entre dos grandes peñascos, una cabeza con capuchón rojo: era Balin que oteaba alrededor.

Bilbo tenía ganas de palmotear y gritar de alegría, pero no lo hizo.

Todavía llevaba puesto el anillo, por miedo de encontrar algo inesperado y desagradable, y vio que Balin estaba mirando directamente hacia él sin verlo.

"Les daré a todos una sorpresa", pensó mientras se metía a gatas entre los arbustos del borde de la cañada.

Gandalf estaba deliberando con los enanos.

Hablaban de todo lo que había ocurrido en los túneles, preguntándose y discutiendo qué irían a hacer ahora.

Los enanos refunfuñaban, y Gandalf decía que de ninguna manera podían continuar el viaje dejando al señor Bolsón en manos de los trasgos, sin tratar de saber si estaba vivo o muerto, y sin tratar de rescatarlo.

Al fin y al cabo es mi amigo dijo Gandalf, y una buena persona.

Me siento responsable.

Ojalá no lo hubieseis perdido.

Los enanos querían saber ante todo por qué razones lo habían traído con ellos, por qué no había podido mantenerse cerca y venir también, y por qué el mago no había elegido a alguien más sensato.

Hasta ahora ha sido una carga de poco provecho dijo uno.

Si tenemos que regresar a esos túneles abominables a, buscarlo, entonces maldito sea, digo yo.

Gandalf contestó enfadado: Lo traje, y no traigo cosas que no sean de provecho.

c) Fragmento de la traducción con MTUOC (Rainbow-ModernMT)

Capítulo 6:

De la pantera en el fuego

Bilbo había escapado a los duendes, pero no sabía dónde estaba.

Había perdido la capucha, el manto, la comida, el poni, sus botones y sus amigos.

Se paseó de un lado a otro, hasta que el sol comenzó a descender hacia el oeste, detrás de las montañas.

— ¡Oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!, ¡oh!,

Luego miró hacia delante y pudo ver antes de que él sólo se deshiciera y las laderas cayeran hacia las tierras bajas y las llanuras bañadas de vez en cuando entre los árboles.

— ¡Cielo santo! — exclamó —.

¡Me parece que tengo derecho al otro lado de las montañas brumosas, justo al borde de la tierra más allá!

¿DÓNDE y en dónde pueden llegar los hermanos Gandalf y los otros?

Yo sólo espero a la bondad que todavía no están allí, en el poder de los duendes.

Seguía vagando, por el pequeño valle, por encima del borde, y por las laderas más allá; pero todo el tiempo un pensamiento muy incómodo crecía dentro de él.

Se preguntó si no debería, ahora que tenía el anillo mágico, volver a los horribles, horribles, túneles y buscar a sus amigos.

Acababa de decidir que era su deber, que debía volver, y que se sentía muy desgraciado por él, cuando oyó voces.

Se detuvo y escuchó.

No se oía el ruido de los duendes; por lo que avanzó cuidadosamente.

Estaba en un sendero pedregoso que se deslizaba hacia abajo con una pared rocosa. Sobre la mano izquierda; al otro lado, el suelo se extendía, y allí estaban las aves por debajo del nivel del camino que colgaba de entre los arbustos y los árboles bajos.

En una de esas aves estaba hablando la gente de los arbustos.

Se arrastró aún más cerca, y de repente vio entre dos grandes rocas una cabeza con una capucha roja sobre: era Balin haciendo el examen.

— ¿Qué es lo que ha hecho? — le preguntó.

Tenía todavía el anillo, por miedo a encontrarse algo inesperado y desagradable, y vio que Balin le miraba fijamente sin fijarse en él.

" Les daré toda una sorpresa", pensó, mientras se arrastraba hacia los arbustos en el borde de la dell.

Gandalf estaba discutiendo con los enanos.

Estaban discutiendo todo lo que les había sucedido en los túneles, preguntándose y debatiendo lo que iban a hacer ahora.

Los habanos estaban gruñendo, y Gandalf estaba diciendo que no podían seguir adelante con su viaje dejando al Sr. Baggins en las manos de los duendes, sin tratar de averiguar si estaba vivo o muerto, y sin tratar de salvarlo.

— Después de todo, es mi amigo — dijo el hechicero —, y no es malo.

Me siento responsable por él.

Deseo la bondad de no haberlo perdido.

Los habanos querían saber por qué había sido llevado, por qué no podía quedarse con sus amigos y venir con ellos, y por qué el hechicero no había elegido a alguien con más sentido.

" Ha sido más problemas que usar hasta ahora", dijo uno.

"Si tenemos que volver ahora a esos tunantes abominables para buscar para él, entonces lo secan, digo. "

— Le he traído, y no traigo las cosas que no sirven de nada.

d) Fragmento de la traducción con Google Translate

Capítulo 6:

Fuera de la sartén al fuego

Bilbo había escapado de los duendes, pero no sabía dónde estaba.

Había perdido capucha, capa, comida, pony, sus botones y sus amigos.

Paseó y siguió, hasta que el sol comenzó a hundirse hacia el oeste, por lo que las montañas.

Sus sombras cayeron por el camino de Bilbo, y él miró hacia atrás.

Luego miró hacia adelante y pudo ver antes que él solo crestas y laderas que caen hacia tierras bajas y llanuras vislumbradas ocasionalmente entre los árboles.

"¡Cielos!" el exclamó.

"¡Parece que he llegado directamente al otro lado de las montañas brumosas, justo hasta el borde de la tierra más allá!

¿Dónde y o dónde pueden llegar Gandalf y los enanos?

¡Solo espero que la bondad todavía no estén en el poder de los duendes! "

Aún de la pequeña valle, sobre su borde, y bajó por las laderas más allá; Pero todo el tiempo un pensamiento muy incómodo estaba creciendo dentro de él.

Se preguntó si no debería, ahora tenía el anillo mágico, para volver a los horribles, horribles, los túneles y buscar a sus amigos.

Acababa de decidir que era su deber, que debía volverse y muy miserable que sintió al respecto, cuando escuchó voces.

Se detuvo y escuchó.

No sonaba como duendes; Entonces se arrastró con cuidado.

Estaba en un camino pedregoso que terminaba hacia abajo con una pared rocosa. en la mano izquierda; En el otro lado, el suelo se dejaba caer y había Dells debajo del nivel del camino cubierto con arbustos y árboles bajos.

En uno de estos Dells debajo de los arbustos, la gente estaba hablando.

Se arrastró aún más cerca, y de repente vio mirar entre dos grandes rocas una cabeza con una capucha roja: estaba haciendo Balin.

Podría haber aplaudido y gritado para alegría, pero no lo hizo.

Todavía había puesto el anillo, por temor a conocer algo inesperado y desagradable, y vio que Balin lo estaba mirando directamente sin notarlo.

"Les daré una sorpresa a todos", pensó, mientras se arrastraba en los arbustos al borde del Dell.

Gandalf estaba discutiendo con los enanos.

Estaban discutiendo todo lo que les había sucedido en los túneles, y preguntándose y debatiendo lo que debían hacer ahora.

Los enanos se quejaban, y Gandalf decía que no podían continuar con su viaje dejando al Sr. Baggins en manos de los Goblins, sin tratar de averiguar si estaba vivo o muerto, y sin tratar de rescatarlo.

"Después de todo, él es mi amigo", dijo el mago, "y no un mal pequeño.

Me siento responsable de él.

Deseo que la bondad no lo habías perdido ".

Los enanos querían saber por qué había sido traído en absoluto, por qué no podía quedarse con sus amigos y acompañarlos, y por qué el mago no había elegido a alguien con más sentido.

"Ha sido más problemas que usar hasta ahora", dijo uno.

"Si tenemos que volver ahora a esos túneles abominables para buscarlo, entonces Dratlo, digo".

Gandalf respondió enojado: "Lo traje y no traigo cosas que no sirven de nada.