TRABAJO DE FIN DE MÁSTER

2021-2022

LA WIKIPEDIA COMO FUENTE DE DATOS PARA UN CORPUS ES-NL

MÁSTER EN TRADUMÁTICA: TECNOLOGÍAS DE LA TRADUCCIÓN

FACULTAD DE TRADUCCIÓN E INTERPRETACIÓN

Autoría

Fernández Veloso, Ana María

TUTOR/A
Ona de Gibert

Barcelona, 26-06-2022

Datos del TFM / *Dissertation data / Dades del TFM*

Título: La Wikipedia como fuente de datos para un corpus paralelo ES-NL

Title: Wikipedia as data source for an ES-NL parallel corpus

Títol: La Wikipedia com una font de dades per un corpus paral·lel ES-NL

Autor/a: Ana Fernández y Fee Dooms

Author: Ana Fernández and Fee Dooms

Tutor: Ona de Gibert

Tutor: Ona de Gibert

Centro: Universidad Autónoma de Barcelona

Centre: Autonomous University of Barcelona (UAB)

Estudios: Máster oficial en Tradumática: Tecnologías de la Traducción

Studies: Official master's degree in Tradumatics: Translation Technologies

Palabras clave / Keywords / *Paraules claus*

Lingüística computacional, corpus paralelo, traducción automática, TA, traducción automática neuronal, TAN, lengua española, lengua neerlandesa, entrenamiento de un motor de traducción, MutNMT

Computational linguistics, Paralel corpus, Machine Translation, MT, Neural Machine Translation, NMT, Spanish Language, Dutch Language, machine translation training, MutNMT

Lingüística computacional, corpus paral·lel, traducció automàtica, TA, traducció automàtica neuronal, TAN, llengua espanyola, llengua neerlandesa, entrenament d'un motor de traducció, MutNMT

Resumen del TFM / *Abstract / Resum del TFM*

En plena era de la globalización, la comunicación internacional se vuelve indispensable en el proceso de evolución económica, política, social y cultural de cada país. En este caso concreto, la barrera terrenal que existe entre España, Bélgica y Países Bajos se ve ahora difuminada por el creciente turismo entre dichas áreas. Y, consecuentemente, la actual situación lingüística de sendas regiones pone en evidencia la necesidad de recursos fiables y de calidad que permitan el acercamiento, la comunicación fluida y el aprendizaje de sus lenguas. Por ello, el presente trabajo propone, por un lado, la construcción de un corpus paralelo neerlandés – español como combinación lingüística, y el entrenamiento de un motor de traducción automática neuronal (TAN) a partir del mismo; y por el otro, su posterior evaluación y comparación con un motor existente ya entrenado.

During these times of globalization, international communication is becoming indispensable in the process of economic, political, social and cultural evolution of each country. In this particular case, the barrier that exists between Spain, Belgium and the Netherlands is now blurred by the growing tourism between these areas. And, consequently, the current linguistic situation of both regions highlights the need for reliable and quality resources that allow the rapprochement, fluid communication and learning of their languages. Therefore, the present thesis proposes, on one hand, the construction of a parallel Dutch-Spanish corpus as a linguistic combination, and the training of a neural machine translation engine (TAN) based on this same corpus; and on the other hand, its evaluation and comparison with an existing engine that had already been trained.

En plena era de la globalització, la comunicació internacional es fa indispensable en el procés d'evolució econòmica, política, social i cultural de cada país. En aquest cas concret, la barrera terrenal que hi ha entre Espanya, Bèlgica i els Països Baixos es veu ara difuminada pel creixent turisme entre aquestes àrees. I, consegüentment, l'actual situació lingüística d'aquestes regions posa en evidència la necessitat de recursos fiables i de qualitat que permetin l'apropament, la comunicació fluida i l'aprenentatge de les llengües. Per això, aquest treball proposa, d'una banda, la construcció d'un corpus paral·lel neerlandès - espanyol com a combinació lingüística, i l'entrenament d'un motor de traducció automàtica neuronal (TAN) a partir d'aquest; i de l'altra, la posterior avaluació i comparació amb un motor existent ja entrenat.

Agradecimientos

El presente trabajo no hubiese sido posible sin la ayuda principal de nuestra tutora, Ona de Gibert, quien nos ha apoyado y guiado durante toda su elaboración. Por otro lado, nos gustaría agradecer la ayuda desinteresada de amigos y familiares, y el apoyo incondicional recibido por parte de nuestras familias.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS Y GRÁFICAS

I. Introducción

1. Presentación y justificación

En plena era de la globalización, la comunicación internacional se vuelve indispensable en el proceso de evolución económica, política, social y cultural de cada país. En este caso concreto, la barrera terrenal que existe entre España, Bélgica y Países Bajos se ve ahora difuminada por el creciente turismo entre dichas áreas. Y, consecuentemente, la actual situación lingüística de sendas regiones pone en evidencia la necesidad de recursos fiables y de calidad que permitan el acercamiento, la comunicación fluida y el aprendizaje de sus lenguas.

De esta manera, el presente trabajo se enmarca en las áreas del estudio de la lingüística computacional, la traducción y la enseñanza de lenguas extranjeras y pretende servir de puente entre las culturas española, belga y holandesa.

Es por ello por lo que se propone llevar a cabo, por un lado, la construcción de un corpus paralelo neerlandés – español como combinación lingüística, y el entrenamiento de un motor de traducción automática (TA) propio a partir del mismo; y por el otro, su posterior evaluación y comparación con un motor existente ya entrenado.

Como primera idea, se pretende elaborar un corpus desde cero a partir de artículos de la Wikipedia en ambas lenguas de trabajo. Se ha escogido esta fuente porque recoge información de calidad que ha sido evaluada previamente, lo que la hace más fiable por lo general que los datos que se puedan obtener directamente de internet.

Una vez creado el corpus paralelo, se procederá al entrenamiento un motor de traducción automática neuronal (TAN) que facilitan desde la universidad. Este motor va especialmente bien, por lo que se cuenta con la ventaja de simplemente tener que introducir y trabajar con datos propios elaborados por el equipo, pudiendo así prescindir de tener que realizar ningún ajuste necesario en la programación del sistema.

En cuanto se haya realizado el entrenamiento con el motor TAN, se llevará a cabo un análisis contrastivo entre aquellos resultados obtenidos del motor elaborado por el equipo y un motor de código libre entrenado a partir de colecciones obtenidas del banco de datos de Opus Corpus, concretamente el corpus de Wikimatrix.

La motivación que lleva a adentrarse en este trabajo es el interés común y personal en los estudios de TA, concretamente los de TAN. Además, el hecho de contar con la ventaja de tener

una integrante de nacionalidad belga en el equipo de trabajo ha supuesto la oportunidad ideal para elaborar una herramienta útil de la que puedan beneficiarse los hablantes de ambas lenguas de estudio.

Como motivo principal, el desconocimiento y los escasos recursos encontrados para esta combinación lingüística ha servido de razón de peso para intentar ofrecer un recurso más y de calidad para investigadores, lingüistas, traductores, estudiantes, y, en definitiva, cualquier apasionado que quiera aprender una de estas dos lenguas.

Por otro lado, y tras una breve labor de documentación, se llega a la conclusión de que aquellos escasos recursos son en su mayoría de carácter privado además de muy especializados (por lo general, de índole jurídica, económica y técnica). Por lo que se precisa aportar una herramienta pública y accesible para el beneficio, uso y disfrute de todos.

2. Objetivos

Con este trabajo se pretende arrojar luz sobre la importancia de la creación de un corpus de estas características, justificar con datos la necesidad de dicho corpus y de un motor de TA y sus múltiples aplicaciones.

El objetivo principal que se persigue es la creación de un motor TAN útil para los hablantes de estas lenguas y entrenado con datos lingüísticos fiables. Así pues, una vez creado el motor, se compararán los resultados de traducción que se obtienen a partir de este con aquellos obtenidos de un motor de traducción ya entrenado a partir de datos no controlados.

Creemos que nuestro motor, que sabemos que estará elaborado a partir de datos de calidad, podrá superar en cuanto a calidad a otros modelos de motores de traducción e intentaremos proporcionar evidencias mediante un control de calidad estadístico para tal fin.

Como conclusión y considerando los objetivos del presente trabajo, nos hemos propuesto responder dos preguntas de investigación al finalizar nuestro proyecto:

1. ¿Es mejor un corpus con una gran cantidad de datos y de mala calidad o un corpus con menos datos, pero corregido y de calidad?
2. ¿Cuál es el mínimo de datos necesarios para poder obtener un corpus de calidad?

Queremos hacer esta comparación para estudiar a qué se debe dar más importancia a la hora de aportar datos lingüísticos a un motor de traducción. Consideramos una parte crucial del

trabajo llegar a una conclusión clara sobre qué motor es mejor, el que se entrena a partir de corpus con una gran cantidad de datos, pero sin haber pasado por un control de calidad, o el que entrenamos nosotros, a base de un corpus más pequeño, pero de calidad lingüística más alta y controlada.

3. Organización del trabajo

Idealmente, nos organizaremos a partir de un calendario, teniendo en cuenta las diferentes entregas del trabajo. Nuestro trabajo se dividirá en distintas partes: la introducción, el marco teórico, la metodología y análisis de los resultados y las conclusiones.

Para la entrega que se hace en marzo, nos gustaría tener hecho todo el marco teórico, es decir, redactar toda la información necesaria para poder llevar a cabo el trabajo. Esto conlleva tener casi bien toda la bibliografía y las fuentes citadas.

Para la entrega que se hace en abril, queremos tener descrita toda la metodología, e idealmente tener nuestro propio corpus hecho.

Por último, en la entrega final, habremos entrenado nuestro motor con los datos del corpus creado y habremos comparado sus resultados con los resultados del otro motor.

El trabajo está organizado como se detalla a continuación: En esta primera sección se exponen los objetivos que se persiguen y se argumenta la motivación que lleva a una investigación de esta tesitura.

La segunda sección corresponde al marco teórico del proyecto. En él se pondrá en contexto la situación lingüística de los territorios principales en los que se habla una de las lenguas de trabajo, esto es, el neerlandés. Seguidamente, explica en profundidad el fenómeno de la traducción automática (TA), desde sus aplicaciones hasta las tipologías que abarca. En este punto, se prestará especial interés al modelo de traducción automática neuronal (TAN), eje del presente trabajo. De esta manera, se explicará minuciosamente su funcionamiento interno, el proceso de entrenamiento y las métricas de evaluación vigentes para analizar los resultados. Por último, se hablará de los corpus en general y de los corpus paralelos en concreto, lo que pondrá fin a un marco teórico que recoge toda la información necesaria para poder pasar a la siguiente sección.

La tercera sección se centra en el marco metodológico del proyecto. En él se llevará a cabo, por un lado, la creación de un corpus paralelo con los datos que se puedan encontrar en los *dumps* de Wikipedia en la combinación español-neerlandés; y por el otro, se realizará un entrenamiento de un motor TAN a través de la plataforma MutNMT a partir de los corpus que han sido creados previamente.

En las siguientes secciones, se expondrán los resultados obtenidos durante todo el proceso, se sacarán unas conclusiones finales en las que se responderán a las preguntas de investigación, y se presentará una extensa bibliografía con los artículos que han servido de guía y apoyo al proyecto.

II. Marco teórico

1. El neerlandés

En primer lugar y, por las razones anteriormente expuestas, es imperativo poner en contexto la situación actual que ocupa la lengua neerlandesa dentro del panorama mundial, y más concretamente su relación con la lengua española. Por otro lado, previo a dicha contextualización, prima clarificar los conflictos que los términos que rodean tal lengua y sus variedades lingüísticas despiertan.

Desde tiempos inmemoriales en toda la historia mundial, los cambios políticos y demográficos dejan mella en la situación lingüística de las regiones en las que se den. Es por ello por lo que la zona que hoy en día ocupan los Países Bajos y Bélgica ha estado expuesta a una innumerable cantidad de alteraciones que justifican la confusión de su realidad geográfica y lingüística.

En 2007, Valembois comentaba la complejidad del asunto al haberse utilizado vocablos relacionados (esto es, «flamenco», «holandés», «neerlandés», entre otros) de manera errónea para abarcar contenidos dispares y dando lugar a malentendidos y aclaraba con evidencias históricas la cuestión. De esta manera, se recogen a continuación los términos tratados en su trabajo junto a una breve definición:

a. Flamenco: es la expresión regional belga del neerlandés que ocupa las regiones de Flandes y gran parte de Bruselas.
b. Holandés: gentilicio y lengua oficial de los Países Bajos.
c. Belga: gentilicio de los habitantes de Bélgica.
d. Holanda: corresponde a las partes del actual territorio costero al oeste dentro de los Países Bajos, con las ciudades de Ámsterdam, Rótterdam y La Haya.
e. Neerlandés: nombre oficial del idioma. Este guarda cierto parecido con el nombre de "País Bajo", significado de «Nederland» en neerlandés. Además, neerlandés es el nombre que recibe la persona que habla la lengua.
f. Dutch: término en inglés y que significa "neerlandés".

Se puede coincidir en que fue en 1648 con el Tratado de Münster que se originó el error principal de considerar el «holandés» y el «flamenco» dos lenguas diferentes cuando realmente son variedades regionales de una misma, el neerlandés, a excepción de ligeras variaciones lexicográficas y entonación. Entrando en más detalles, dicha problemática se dio en Bélgica, en

la que nació el término «flamenco» en bocas francófonas para identificar despectivamente el neerlandés que se habla en la parte norte del país. (Valembois, 2007)

Dicho esto, para una mayor claridad y a modo de resumen, se cita textualmente a Valembois (2007):

> *[…]unos seis millones de belgas de pasaporte viven en Bélgica, constituyendo el 60% de su población; en el caso de los Países Bajos, unos quince millones de neerlandeses viven en su comunidad nacional, pero ambos grupos, conformando los millones citados de flamencos y neerlandeses, respectivamente, viven, piensan y aman en neerlandés.*

Desde una panorámica mundial de la lengua, cabe mencionar que el neerlandés se habla en poblaciones de 150.000 y 60.000 habitantes de Francia (Flandes francés) y Alemania, respectivamente. Además, es la lengua oficial de la República de Surinam y de las Antillas Holandesas (Curazao, San Martín, Aruba, Bonaire, Saba y San Eustaquio), así como 5 millones de personas la toman como lengua materna a través de la variante afrikáans, desarrollada por los primeros colonos que llegaron a Sudáfrica en el siglo XVII. (Valembois, 2007) Por lo tanto, aunque parezca una lengua minoritaria en Europa, un número considerable de personas en todo el mundo la habla. (Donaldson, 1983; Valembois, 2007) Tanto es así que el neerlandés forma parte de los once idiomas oficiales de la Unión Europea, ocupando el sexto puesto en importancia dentro del organismo después del inglés, alemán, francés, español e italiano, en ese orden. (Valembois, 2007)

1.1 Territorio de Bélgica

Llegado este punto, conviene resolver en última instancia un entuerto lingüístico que concierne al territorio geográfico belga.

Este país se consagra como tal en 1830, «formalizando también la separación de un matrimonio forzado con los Países Bajos» (Valembois, 2007) después de convivir bajo el reinado de los Habsburgo españoles y austríacos. Hoy en día se mantiene dicha separación, además de otra interna que establece una frontera lingüística desde 1962 para separar la zona de habla neerlandesa en el norte, de aquella de habla francesa al sur, a excepción de Bruselas capital, única región oficialmente bilingüe. (Valembois, 2007; Niederländische Philologie, FU Berlin, 2022)

Así pues, hablamos de un territorio con cuatro áreas lingüísticas y tres idiomas oficiales presentes en los siguientes porcentajes presentados por Vermeiren: «el neerlandés (hablado por el 60% de la población), el francés (39%) y el alemán (1%)». (2016)

1.2 Relación neerlandesa-española

Sin más rodeos, abarcamos de cerca la relación que mantienen los territorios que atañen a este estudio. Y es que a lo largo de la historia se ha dado una sucesión de sorprendentes coincidencias históricas que pone en evidencia los lazos que hoy en día guardan los Países Bajos y Bélgica con España.

Como hemos mencionado anteriormente, los antiguos «Países Bajos» así como las colonias españolas estuvieron casi dos siglos bajo el mismo estado. Fue una «vinculación forzada, de política dinástica, pero su duración e intensidad dejaron desde luego huellas por ambos lados». (Valembois, 2007) Prueba de ello son los escritos de reconocidos autores del Siglo de Oro, pues Lope de Vega y Calderón de la Barca haciendo referencia en varias ocasiones a la faena española por la zona de Flandes. Otro caso de interferencia lo encontramos en vocablos provenientes del neerlandés, como por ejemplo «loterij» para «lotería». (Valembois, 2007)

No obstante, la relación existente hoy en día continúa siendo cercana, pero se aleja de reinados comunes y referencias literarias para acercarse al plano idiomático y turístico.

Respecto de lo primero, Vermeiren (2016) nos explica que «el inglés, el francés y el alemán están más presentes como lengua de estudio en Bélgica y Holanda debido a su mayor peso demográfico». En cuanto al italiano y al español, al ser lenguas menos centrales, se estudian menos. Esto se debe, según Pomar (2006), a factores políticos, legislativos y sociales, principalmente. Tanto es así que ambos países cuentan con exhaustivos programas de incorporación del extranjero «como conciencia propia en consideración de la necesaria preparación frente a la interferencia de lenguas». (Valembois, 2007) De este modo, en niveles de posgrado se llega al dominio de varias lenguas, de entre las que destacan en primer lugar las cooficiales en el caso de Bélgica, seguidas del español. (Pomar, 2006; Valembois, 2007)

Respecto de lo segundo, estos datos se apoyan directamente en la situación del turismo entre los países en cuestión. En este sentido, España encabeza la lista de destinos turísticos a nivel europeo, y está entre las primeras posiciones a nivel mundial. Más concretamente, los últimos informes sobre competencia turística revelan que los países de los que proceden un mayor número de turistas gracias a nuestro llamativo segmento de «sol y playa» son Reino Unido, Alemania, Francia, Italia y Países Bajos. (Gómez y González, 2014; Peña, 2017)

Con todo, se puede confirmar la latente relación entre sendos países y se pone de manifiesto la imperiosa necesidad de bibliografía y datos de referencia que ayuden a forjar un acercamiento lingüístico de calidad entre la lengua neerlandesa y española.

2. La traducción automática (TA)

Después de presentar a grandes rasgos el neerlandés, su contexto, y todo lo relativo al idioma pertinente para este trabajo, en este nuevo apartado se pretende abordar el fenómeno de la traducción automática.

Son muchos los que se han aventurado desde su nacimiento a dar una definición exacta del concepto. A continuación, se ofrecen dos aportaciones de parte de expertos de la materia, Berner (2003) y Forcada (2010), respectivamente, cuyos trabajos se han utilizado para el desarrollo del presente:

> *Machine translation (MT) is the use of computer software to translate text or speech from one natural language into another. Like translation done by humans, MT does not simply involve substituting words in one language for another, but the application of complex linguistic knowledge: morphology (how words are built from smaller units of meaning), syntax (grammar), semantics (meaning), and understanding of concepts such as ambiguity.*
>
> *Machine translation (MT) is the translation, by means of a computer using suitable software, of a text written in the source language (SL) which produces another text in the target language (TL) which may be called its raw translation».*

Los orígenes de la traducción automática los podemos encontrar en Francia y la URSS (1933-1945), donde se crearon los primeros sistemas de TA de la mano de George Artsrouni y Petr Troyanskii. Ambos sistemas empezaron como diccionarios bilingües automatizados, pero solo la propuesta de Troyanskii supuso un avance significativo al llegar a incorporar una memoria y componentes electrónicos. Sin embargo, ya existía el ordenador para cuando se dio a conocer su idea años después de su invención. (Hutchins, 2009, 2014)

Pronto se puso de manifiesto la complejidad de las reglas sintácticas en las que se basaba el sistema de diccionarios, haciendo así evidente la necesidad de encontrar otros medios de análisis sintácticos más sistemáticos. Por consiguiente, surgieron más adelante diferentes proyectos inspirados en modelos de la gramática formal pero, desafortunadamente, también estos presentaron problemas de índole semántica para los que los investigadores no consiguieron encontrar solución. (Hutchins, 2014)

Fue entonces cuando, ante la ausencia de avances en la materia, en EE. UU. se forma el comité ALPAC (Automatic Language Processing Advisory Committee) que concluyó en la redacción de un informe el cual determina que la TA es lenta, poco exacta y el doble de cara que la traducción humana, y da su negativa a la financiación para su investigación y desarrollo. (Hutchins, 2014)

En el tiempo comprendido entre 1966 y 1990 emergieron nuevos sistemas y se desarrollaron técnicas más avanzadas por parte de países como Canadá, Alemania y Francia que dejaron atrás el predominio exclusivo del enfoque basado en reglas. No obstante, no fue hasta finales de la época que se llegó a un punto de inflexión con el surgimiento de una amplia variedad de herramientas integradas en la labor traductora (originalmente, «*translator's workbench*»). (Hutchins, 2007, 2014) Entre estas, tal y como lista Hutchins (2007), se encuentran «multilingual word processing, OCR facilities, terminology management software, facilities for concordancing, and in particular "translation memories".»

Un último avance de principios de 1990 fue el aumento de la actividad de la TA para la investigación sobre sus aplicaciones prácticas, el desarrollo de puestos de trabajo para traductores profesionales, el funcionamiento de sistemas de lenguaje controlado y dominio restringido y la aplicación de funciones de traducción en sistemas de información multilingües. (Hutchins, 2014)

Este breve paso por la historia de la TA muestra únicamente los inicios de lo que es hoy en día un fenómeno en pleno auge, tanto en volumen, como en medios de aplicación. Cada vez son más las empresas a nivel mundial que incorporan servicios de traducción, ya sea para el preprocesamiento de entradas o como posedición. Asimismo, el acceso a la TA se ha generalizado de manera que el público puede hacer uso gratuito de servicios de TA en línea, dejando así atrás los limitados softwares de PC. (Hutchins, 2014)

Pero lejos de ser esto una amenaza, como en estos días se especula, Hutchins justifica a la perfección la convivencia del traductor profesional y el uso de su «enemigo» la TA al decir que los traductores profesionales hacen uso de la TA así como de las memorias de traducción como herramientas en la producción de borradores de traducción. (Hutchins, 2014) Esta afirmación es en la actualidad un debate abierto que ha incitado a los traductores a posicionarse a favor o en contra de la TA en la profesión, pero eso supone un caso que se desvincula de la razón que se persigue en el presente trabajo.

2.1 Aplicaciones de la traducción automática

El fenómeno de la traducción automática recoge múltiples clasificaciones, aun así, las que en este punto se estudiarán aquellas desde el punto de vista de su aplicación y de su arquitectura.

Primeramente, se puede afirmar que la traducción automática tiene varias formas de uso según la finalidad que se quiera conseguir:

Comprensión o asimilación (*assimilation*): es el que ofrece una traducción en crudo que te permite sacar una idea general del contenido sin necesidad de conocer la lengua de origen, ya que lo importante es transmitir el sentido del mensaje en un contexto. A este uso también se le llama *Machine Translation for Watcher* y fue la aplicación inicial con el que se empezó a conocer la TA. Un ejemplo de este sería la traducción automática de páginas web para una comprensión general de lo que se ofrece, o la traducción de chats y foros. (Ginestí y Forcada, 2009; Chéragui, 2009)

Publicación o diseminación (*dissemination*): es el uso de la traducción automática como paso intermedio en la producción final de un texto en la lengua de destino que se publicará en algún momento. Este texto tendrá que ser poseditado por un profesional para resolver cualquier ambigüedad o cambio de sentido que la traducción automática no percibe. Se le conoce también como *Machine Translation for Reviser* (Ginestí y Forcada, 2009; Chéragui, 2009)

Sin embargo, en ninguno de los dos usos se concebía la obligatoria implicación del traductor, pues, tal y como expone Hutchins, «MT was not seen as a computer aid for translators» (2009), sino como una funcionalidad aparte.

2.2 Tipos de traducción automática

Desde el punto de vista de su arquitectura, en este caso se distingue entre arquitectura lingüística y arquitectura computacional. La primera se presenta con la siguiente figura, el triángulo de Vaoquois, en la que se muestra una interrelación entre todas sus construcciones (Forcada, 2010; López, 2018):

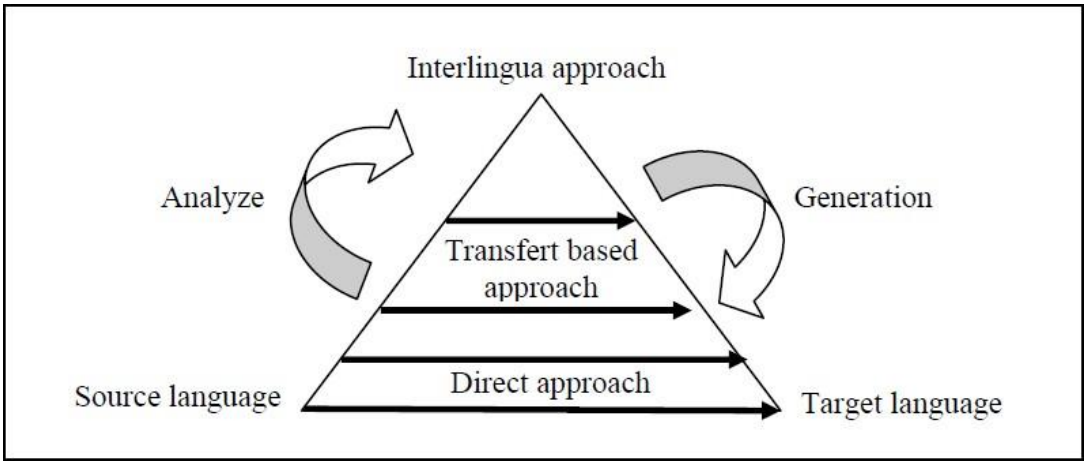


Figura 1: Triángulo de Vauquois (1968)

En cuanto a la arquitectura computacional, su clasificación en diferentes motores de traducción automática es hoy la más conocida. Así, destacan los siguientes:

2.2.1 La traducción automática basada en reglas

Desde 1990, el enfoque basado en corpus es el que ha estado en el punto de mira y ha sido centro de los principales avances mientras que, en segundo plano, se continuaban las investigaciones en el campo de la TA basada en reglas. (Hutchins, 2010)

Integrados en este último, los expertos diferencian tres sistemas, cada uno de ellos con sus particularidades, pero estrechamente enlazados. Así pues, para situarlos en contexto de una manera clara y visual, se expone a continuación un esquema general de la traducción automática basada en reglas en el que se definirán todos sus términos.

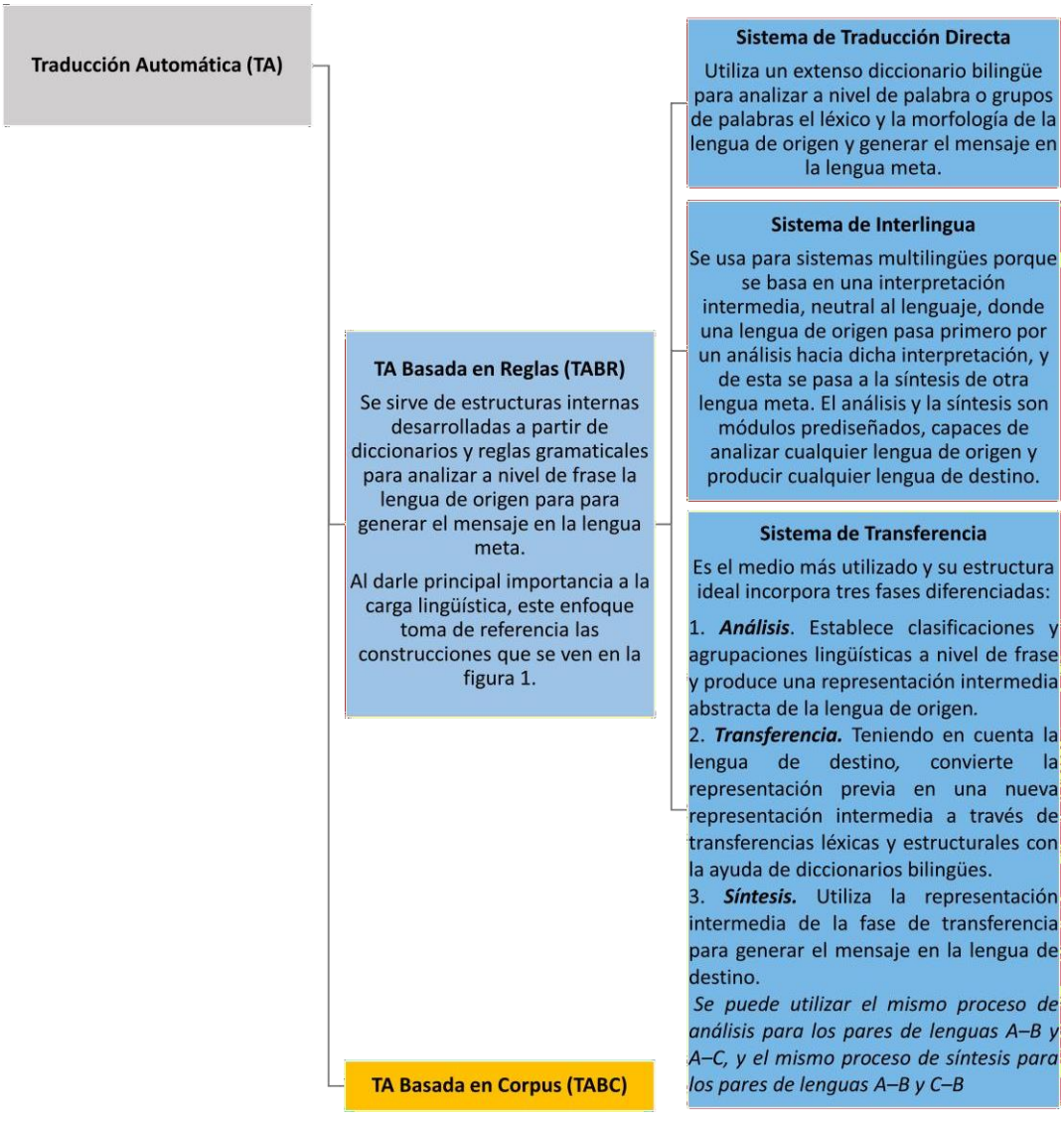


Figura 2: Esquema sobre la traducción automática basada en reglas

2.2.2 La traducción automática basada en corpus

El renacimiento del enfoque basado en corpus, en palabras de Hutchins, fue un acontecimiento que se tomó «como una vuelta al "empirismo" de la primera década, así como un reto para el "racionalismo" del enfoque basado en reglas, predominante en los 70 y 80». El motivo de ello es una nueva perspectiva de la estadística, la cual en un principio estaba delimitada por reglas lingüísticas, pero que ahora se ciñe a los principios estadísticos de lo estrictamente analítico.

De nuevo, se incorpora a continuación un esquema que entrará en detalle en lo que a definiciones y tipos de TA basada en corpus refiere.

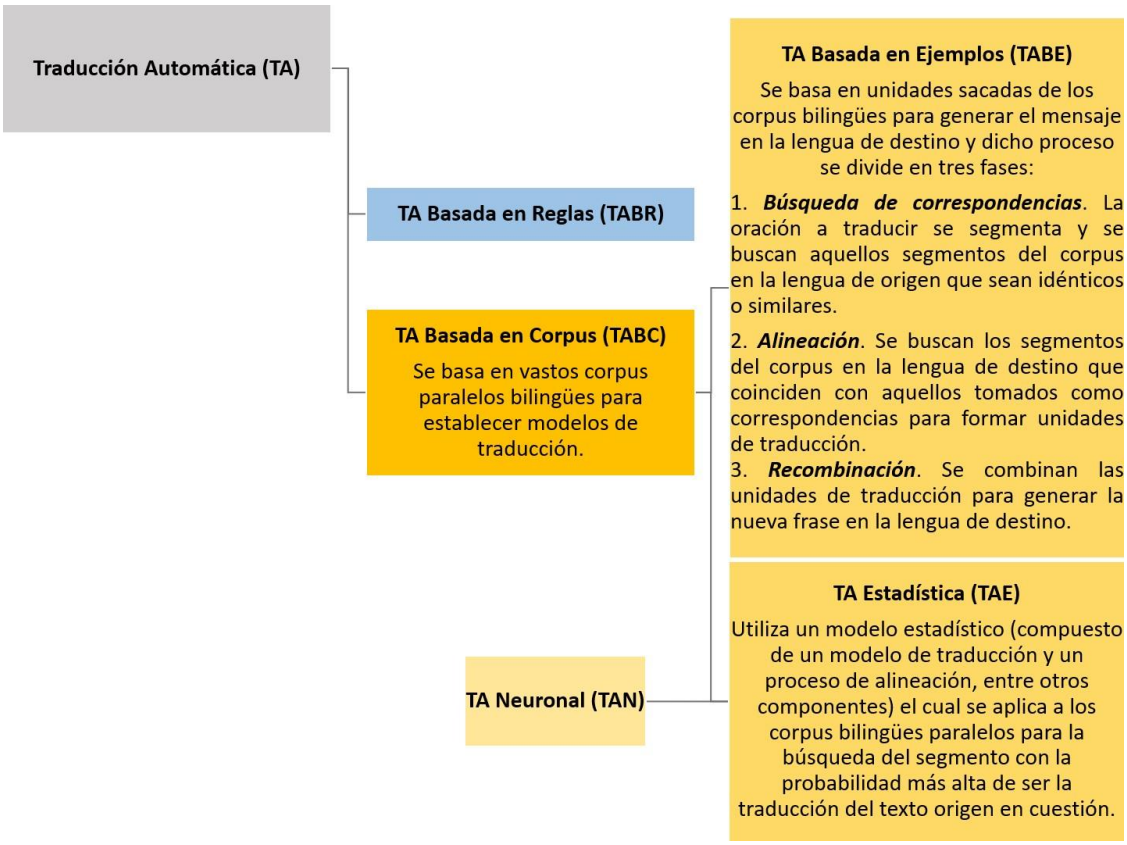


Figura 3: Esquema sobre la traducción automática basada en corpus

Se puede afirmar que la TAE es el marco dominante de la investigación dentro del campo de la traducción automática. Hutchins (2014) recoge las siguientes razones:

> *[…] i) the availability of large monolingual and bilingual corpora, ii) the open-source availability of software for performing basic SMT processes (alignment, filtering, reordering), such as Moses, GIZA, etc.; iii) the availability of widely accepted metrics for evaluating systems (BLEU, and successors).*

Pese a existir un claro predominio del enfoque basado en corpus, hay ciertos procesos de la traducción automática basada en reglas que mantienen una relevancia importante. Estos son: el análisis sintáctico para una mejora en el reordenamiento de las frases entre lenguas de características dispares, por ejemplo, el inglés y el japonés); el tratamiento de lenguas con gran riqueza morfológica, como el ruso o el finés; los problemas de transcripción de los nombres en ciertas lenguas, particularmente en el chino; y los problemas discursivos entre lenguas, como en el caso del tratamiento de pronombres. (2014)

2.2.3 Enfoques híbridos

Dicho eso, en un afán por encontrar la armonía perfecta entre lo mejor de cada uno de los dos enfoques, muchos investigadores han optado por nuevos enfoques «híbridos». Estos nacen de la conciliación entre una mejora de las metodologías de la década anterior y la introducción de nuevas aplicaciones de los procesos de traducción automática, pues hoy se ha llegado a la conclusión de que no es posible conseguir una traducción automática de buena calidad utilizando un único método. (Hutchins, 2010, 2014)

Uno de los modelos híbridos para el que se han llevado a cabo más investigaciones es el sistema *multi-engine*, el cual combina una arquitectura basada en reglas (para un análisis morfológico y sintáctico) y otra basada en corpus (ya sea TAE o TABC) y utiliza métodos estadísticos para analizar los resultados ofrecidos por ambos sistemas y así seleccionar el mejor. (Hutchins, 2009, 2010)

3. La traducción automática neuronal (TAN)

Desde hace algunos años se ha centrado la mirada en un nuevo modelo de traducción automática, la traducción automática neuronal (TAN, por sus siglas; o NMT, del inglés, *Neural Machine Translation*). Mientras guarda ciertas similitudes con su predecesor, la TAN «ha logrado el esperado salto cualitativo» (Pym y Torres-Simón, 2021) y está empezando a tomarle el revelo a la TAE siendo hoy sujeto de la mayoría de las investigaciones en el campo. (Forcada, 2010; Casacuberta y Peris, 2017)

«El adjetivo “neuronal” es puramente metafórico: los sistemas de traducción se siguen basando en métodos estadísticos y en el *deep learning*» (Pym y Torres-Simón, 2021), esto es, aprendizaje profundo. En este caso, el modelo utiliza un enfoque computacional que trabaja con redes neuronales artificiales (*artificial neural networks*) compuesto por capas de neuronas recurrentes

que se retroalimentan entre ellas por lo que se consigue crear temporalidad, permitiendo a la red que tenga memoria. (Andújar, 2021) «Existen las capas de entrada, las de salida y las que están ocultas» y cada una realiza una etapa de computación. (García, 2019)

Un definición más concisa y clara la aportan Castilho *et al.* (2017):

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

En cuanto a su funcionamiento interno, «la activación de cada neurona artificial de la red (su excitación o depresión) depende de la activación de las otras neuronas y del peso de las conexiones. Así, el signo y la magnitud del peso de la conexión determina el comportamiento de la red neuronal». (García, 2019) Si el peso es positivo, las neuronas tienen a excitarse; y si el peso es negativo, las neuronas tienen a deprimirse, aumentando este efecto dependiendo de la magnitud del peso de dichas conexiones. (García, 2019)

Los niveles de activación conjuntos de una capa de neuronas forman lo que se llaman *word embeddings*, esto es, representaciones vectoriales en valores de entre - 1 y + 1 de la información que procesan y que aprenden automáticamente de los grandes corpus en que se basan. (Forcada, 2010; Casacuberta y Peris, 2017; Forcada, Pérez y Rico, 2018; García, 2019)

3.1 Arquitectura codificador-descodificador

En la mayoría de las arquitecturas TAN, el proceso de la traducción lleva intrínseco la codificación de una oración en la lengua de origen y su posterior decodificación en la lengua de origen. (Forcada, 2017)

Tanto el codificador como el decodificador son «una red neuronal (recurrente) artificial especializada con una o varias capas» (García, 2019). En el caso de la red codificadora, esta «se encarga de mapear los *embeddings* de entrada de izquierda a derecha y de derecha a izquierda para generar una representación vectorial de la misma». (Casacuberta y Peris, 2017; Andújar, 2021), la cual está constituida a su vez de representaciones de cada una de las palabras que la forman. (García, 2019)

Tras este proceso, «el decodificador genera la frase destino condicionado por la frase origen», (Casacuberta y Peris, 2017) de manera que, tomando como referencia la palabra previamente generada y el estado de la red neuronal en el momento anterior, predice palabra por palabra escogiendo aquella que tenga la probabilidad más alta de ser la traducción hasta conseguir la

frase destino completa. (Casacuberta y Peris, 2017, García, 2019). Cabe mencionar que, en dicho proceso de selección, entran en juego las traducciones existentes en los corpus, por lo que el uso de una palabra en concreto y su combinación se basan en la coocurrencia estadística recursiva. (Pym y Torres-Simón, 2021)

En última instancia, es conveniente aportar las conclusiones de Pym y Torres-Simón (2021), quienes aseguran que, gracias a este proceso de codificación y descodificación, «[…] la traducción automática neuronal puede proponer soluciones que no corresponderían a una traducción literal. Incluso, es capaz de realizar omisiones estratégicas».

No obstante, «one should never expect the MT system – however good – to understand the text, to always solve ambiguities properly and to produce texts conforming to the TL norms or fit for the intended purpose of the translation.» (Forcada, 2015) Por lo que siempre se necesitará de un profesional que lleve a cabo tareas de posedición para naturalizar el lenguaje y adecuarlo completamente a la lengua de destino.

4. Entrenamiento de un motor TAN

Todos los motores de TA dependen en gran parte de información. Ya sea un motor de TA basada en reglas, como en TA basada en corpus, se pueden distinguir tres componentes principales, según expone Forcada (2015):

> *First, an engine, the program that performs the translation (also called decoder in statistical machine translation); second, the data (either linguistic data or parallel corpora) needed for that particular language pair, and, third, optionally, tools to maintain these data and turn them into a format which is suitable for the engine to use.*

En el caso de los motores TAE, los corpus paralelos se utilizan para entrenar una tabla de traducción estadística compuesta por unidades de traducción de sub-frases (es decir, pares de frases en lenguaje estadístico) e información sobre la probabilidad asociada a cada par. (Forcada, 2015)

Teniendo en cuenta su naturaleza, los sistemas de TA basada en corpus requieren de corpus paralelos alineados a nivel de frase que sean de máxima calidad y en grandes cantidades (en el caso concreto de la TAN, se necesitan miles y miles de pares). (Forcada, 2015; García, 2019) Esto significa que el entrenamiento con esos corpus previamente preparados es un paso imprescindible antes de pasar al proceso de traducción. (Forcada, 2015)

Ahora bien, antes de proceder al entrenamiento de un motor TAN se debe contar con 3 corpus diferentes: *training set* o corpus de entrenamiento, *validation set* o corpus de desarrollo, y *set test* o corpus de test. El primero es el de mayor volumen y es el que se utiliza principalmente para el entrenamiento del motor; el segundo tiene como función detener el entrenamiento a tiempo para que no se "sobreentrene", esto es, para que no aprenda de las traducciones hasta el punto en que le impida crear nuevas composiciones; y por último, el corpus de test, formado por pares de frases que no han sido utilizadas en el entrenamiento y que producen un indicador de rendimiento del sistema completo. (García, 2019)

«Una vez ya preparado el conjunto de datos y creados los modelos de traducción automática, el paso a seguir es entrenar tales modelos. Para ello, es necesaria la especificación de los parámetros asociados a los sistemas de traducción.» (Andújar, 2021) Estos definen la lengua origen, la de destino, la ubicación de los archivos, etc. (García, 2019)

En resumidas cuentas, el funcionamiento interno del programa se puede explicar, en palabras de Forcada (2017), como sigue:

> *We want the neural network to read each source sentence to form distributed representations (values of activations of groups of neurons), such that outputs computed from them are as close as possible to the corresponding reference or gold standard translations in the training set (ideally produced by translation professionals). To that end, one trains the neural network; that is, determines the weight or strength of each of the connections between neurons so that the desired results are obtained.*

Finalmente, se debe tener en cuenta las consideraciones que comenta Forcada en su artículo donde explica la complejidad que supone el entrenamiento de los sistemas TAN ya que requieren de cantidades ingentes de datos a los que traductores profesionales ni pequeñas empresas suelen tener acceso. Además, todo el proceso supone un coste muy elevado, y se debe de contar con *hadwares* específicos que trabajen con potentes unidades de procesamiento gráfico para que los tiempos de entrenamiento se puedan reducir lo máximo posible. (Forcada, 2017; López, 2018)

5. Métricas de evaluación de la TA

La evaluación es todavía un tema abierto y se ha convertido en un campo de vigorosa actividad en cuanto a investigaciones, pues se ha demostrado tener significativa implicación asimismo en

la evaluación de otras áreas, como la lingüística computacional y el lenguaje natural. (Forcada, 2010; Hutchins, 2010)

No obstante, desarrollar la propia definición de calidad supone todo un reto ya que habría que tener en cuenta la intención y finalidad para la que se debe preparar la traducción automática en bruto. En un principio, las medidas de calidad de la traducción automática eran lo que ahora recibe el nombre de evaluación manual o *manual evaluation*. Estas se realizaban por parte de jueces humanos los cuales puntuaban los textos producidos en base a dos criterios independientes: por un lado, la inteligibilidad o fluided (*intelligibility or fluency*) del texto generado por la TA; y, por el otro lado, su precisión o adecuación (*fidelity or adequacy*), esto es, cuanto del significado original se transmite. (Forcada, 2010; Hutchins, 2010)

Aun así, e independientemente de la calidad, valga la redundancia, de tales métodos de evaluación de la calidad, los costes en tiempo y esfuerzo llevaron a desarrollar métodos automáticos desde entrados en el año 2000 en busca de un indicador más preciso. (Hutchins, 2010) Estos, según explica Forcada (2010), «try to measure how close each raw machine-translated sentence is to one or more reference human translations.»

En lo que respecta a métricas de evaluación automática aplicadas en la traducción automática neural, se cita de nuevo a Forcada (2017), quien detalla su funcionamiento como sigue:

> *[TAN automatic evaluation measures] compare the output of the machine translation of usually a single independent professional translation called a reference translation using text similarity measures such as the fraction of matching one-, two, three- and four-word sequences.*

Como único aspecto negativo con relación a la TAN, se ha demostrado que las métricas de evaluación pierden calidad a medida que las frases son más largas. Por otro lado, los últimos informes coinciden en una mejoría general de los resultados de la TAN en contraste con la TAE, concretamente ofrece una disminución de errores en el orden de las palabras, errores morfológicos y errores léxicos. (Forcada, 2017)

Las métricas propuestas que se alzaron más relevantes y que perduran en la actualidad son las siguientes:

5.1 BLEU

El método de evaluación automática más conocido es el de BLEU (BiLingual Evaluation Understudy), propuesto por Papineni en 2001 (Papineni *et al.,* 2001). Esta métrica trabaja con n-gramas, es decir, secuencias de una a cuatro palabras sacadas de la traducción en bruto, y con

traducciones humanas de referencia. (Forcada, 2017; López, 2018) Una definición bastante precisa la aportan Kit y Wong (2015), y dice así: «It is based on counting the number of n-grams, namely sequences of consecutive word(s) of varying length, co-occurring in an MT output and in one or more versions of corresponding reference, usually each in the form of a sentence». En cuanto a los resultados, la puntuación final va en una escala del 0 al 1, pero se suele multiplicar por 100 para una mayor capacidad de interpretación (Andújar, 2021). Por otro lado, Papineni *et al.* (2001) hacen una especificación, y es que lograr un 1 es prácticamente imposible incluso para un traductor humano ya que tendría que la traducción tendría que ser idéntica a la de referencia.

5.2 NIST

En 2005 nace NIST (National Institute of Standards) como una mejora de su predecesor BLEU, por lo que ambos trabajan con traducciones humanas de referencia. En este caso, hay dos diferencias claras entre ambos sistemas: por un lado, en lugar de ponderar cada n-grama por igual como sucede con BLEU, NIST otorga más peso a los n-gramas que aparecen menos, los cuales considera más informativos; y, por otro lado, en lugar de hacer la media geométrica de sus resultados, NIST utiliza la media aritmética. Finalmente, otra modificación que incluye esta métrica es la de la penalización por brevedad para las variaciones más pequeñas en la longitud de las frases traducidas. (López, 2017; C. Kit and T. Wong, 2015)

5.3 METEOR

En 2005 también, Banerjee y Lavie proponen esta métrica de evaluación orientada en el *recall* (la proporción de n-gramas que concuerdan con el número total de n-gramas de referencia). Se cita de nuevo a Kit y Wong (2015), quienes detallan su funcionamiento:

> *It begins with an explicit word-to-word alignment to match every word (i.e., a unigram) in a candidate with a corresponding one, if any, in a reference. To maximize the possibility of matching, it uses three word-mapping criteria: (1) exact character sequences, (2) identical stem forms of word, and (3) synonyms.*

METEOR también aplica una penalización por fragmentación para cada frase con un orden de palabras erróneo y se caracteriza por su gran flexibilidad a la hora de ponderar los parámetros. (Kit y Wong, 2015; López, 2018) Igualmente, este sistema sigue actualizándose y perfeccionando sus funciones, como es el caso de la versión METEOR 3.1. (Denkowski y Lavie, 2011)

Tantos son los avances en este campo, que en 2020 se ha llegado incluso a desarrollar un marco neuronal para entrenar modelos de evaluación de TA multilingües bajo el nombre de COMET. Sus autores (Rei, Stewart, Farinha y Lavie, 2020) lo definen así:

> *COMET […] takes advantage of recent breakthroughs in cross-lingual language modeling to generate prediction estimates of human judgments such as Direct Assessments (DA), Human-mediated Translation Edit Rate (HTER) and metrics compliant with the Multidimensional Quality Metric framework.*

En una línea independiente, aparecen métodos de evaluación que, a diferencia de los sistemas mencionados con anterioridad los cuales se basan en cálculos a partir de los aciertos de la TA, estos miden los errores generados por la TA. Los más importantes son los sistemas de Word Error Rate (WER) y Translation Error Rate (TER). El primero, compara el texto de la traducción automática con su traducción de referencias y calcula el número de ediciones realizadas. El segundo, calcula el número de ediciones que se han de hacer sobre la traducción automática para que concuerde exactamente con su traducción de referencia. (Chéragui, 2009; López, 2018)

Dicho esto, teniendo en cuenta los beneficios y limitaciones de cada métrica, es evidente la necesidad de la presencia de la figura del traductor profesional, bien sea realizando la tarea traductológica en sí, o supervisando y ajustando los parámetros de las métricas estudiadas para mejorar los resultados a la lengua de destino.

6. Los corpus: características y tipologías

Con relación a los corpus, se puede decir que constituye una herramienta multidisciplinaria aplicable a diversas líneas de investigación. De hecho, un corpus lingüístico se define como un conjunto de textos suficientemente extensos en amplitud profundidad que han sido preparados y adaptados para su procesamiento con el fin de ayudar a la investigación lingüística. (Martínez, 2003)

Todo corpus debe reunir tres características principales: la primera es que este debe ser una herramienta representativa «de la variedad u objeto lingüístico» (Martínez, 2003) que se pretende analizar; para ello, ha de contar con una segunda característica que es que «el corpus ha de ser extenso, y buscar variedad de estilos y de registros»; finalmente, los corpus lingüísticos han de estar preparados para procesar «cantidades ingentes de información.» (Martínez, 2003)

Por otro lado, y contando con unas décadas a las espaldas de investigación en el campo a día de hoy no existe una clasificación homogénea de los tipos de corpus. Sin embargo, lo cierto es que muchos expertos se han aventurado a proponer múltiples clasificaciones teniendo en cuenta distintos criterios. Entre los que la mayoría consideró, se pueden encontrar subdivisiones basadas en el origen de los autores de los textos que contiene el corpus (corpus de variedades regionales), o en el tratamiento gramatical que haya recibido el corpus (corpus etiquetados o

no etiquetados), o en el origen y características de los textos que componen el propio corpus (corpus sincrónicos, diacrónicos, escritos, orales, abiertos, cerrados, especializados, genéricos, ect). (Martínez, 2003)

Pero si se tiene en cuenta las lenguas en las que están escritas esos textos, se genera una clasificación en la que se encuentran los tipos de corpus que conciernen en el presente trabajo. Y es que un corpus lingüístico no contiene necesariamente textos en la misma lengua, estos pueden ser monolingües, bilingües y multilingües. (Hallebeek, 1999; Martínez, 2003)

6.1 Los corpus paralelos

Ni que decir tiene que, al ser un campo relativamente nuevo, y al no existir clasificación oficial sobre los tipos de corpus lingüísticos, el estudio de la traducción basada en corpus no ha afianzado siquiera una consistencia de términos. Los mismos investigadores utilizan indistintamente los términos «translation corpus, parallel corpus and comparable corpus» para referirse a «types of cross-linguistic texts.» (Li, 2015) Aun así, se puede decir que la terminología comúnmente aceptada ahora (Hallebeek, 1999; Doval, 2017) hace la siguiente distinción, en palabras de Lan Li (2015):

- *Comparable Corpora: consisting of original texts in two or more languages, matched by criteria such as the time of composition, text category, intended audience, etc.*
- *Translation corpora: consisting of original texts in one language and their translations into one or more other languages.*

El término que falta por definir es el de *parallel corpus* y es el más controvertido de los tres ya que, como se adelantaba al comienzo del capítulo, muchos autores utilizan este término como sinónimo de los dos anteriores ya descritos. No obstante, con el creciente estudio de la TABC, los corpus paralelos se llegan a entender también, en el sentido más amplio, como cualquier colección de textos en diferentes idiomas y variedades que contienen información similar recogida en condiciones pragmáticas similares. Es decir, los corpus paralelos es el término general para recoger a los corpus translingüísticos. (Li, 2015)

Teniendo en cuenta el juego de lenguas, los corpus paralelos pueden ser, por un lado, bilingües o multilingües; y por el otro lado, unidireccionales, bidireccionales o multidireccionales. (Hallebeek, 1999; Doval, 2017)

En cuanto a su utilidad, los corpus paralelos encierran un amplio abanico de aplicaciones multilingües, convirtiéndose así en un recurso indispensable en múltiples campos de

investigación. Entre estos destaca su aplicación en una amplia «gama de tareas de procesamiento de lenguaje natural, como extracción de información y terminología multilingüe y, especialmente, dentro de la traducción automática.» (Doval, 2017) Como señalan Kit y Nie «The training of statistical MT models critically depends on the availability of a large amount of bilingual parallel texts (or bitexts).» (2015)

Es de indiscutible interés mencionar la crucial importancia que ha supuesto el avance de la tecnología en general para todos los campos, y la llegada de internet para el traductor profesional en concreto. (Gaspari, 2015) En la web se pueden encontrar en la actualidad desde metadiccionarios, hasta base de datos terminológicas y glosarios en línea. Tanto es así, que con la llegada del siglo XXI surgen los primeros repositorios de memorias de traducción y corpus paralelos de libre acceso. (Gaspari, 2015; Doval, 2017) «Los corpus paralelos más importantes tanto por su tamaño como por su difusión son los que se derivan de textos producidos en diferentes organismos de la Unión Europea», (Doval, 2017) y se presentan a continuación algunos ejemplos con fines ilustrativos:

En primer lugar, cabe mencionar el JRC-Aquis, primer corpus preprocesado y alineado a nivel de frase que fue distribuido por la Comisión Europea. Su colección multilingüe, que cuenta con la versión 3.0, está basada en textos legislativos de la UE a partir de los años 50 en todas las lenguas europeas oficiales. (Gaspari, 2015; Doval, 2017)

En segundo lugar, nombrar la colección de textos paralelos multilingües de la DGT que se publica en línea como parte de la comisión para apoyar la diversidad lingüística y la reutilización de la información generada por las instituciones de la UE. Este tipo de corpus abarca diferentes aplicaciones, tales como «training statistical MT systems, extracting monolingual or multilingual lexical and semantic resources, developing language models for data-driven systems with a linguistic or translation component, etc.» (Gaspari, 2015)

En tercer lugar, Koehn publicó en 2005 la colección de textos paralelos alineados EuroParl. Esta la compone una recopilación en 11 lenguas de las actas de los plenos del Parlamente desde 1996. (Doval, 2017)

En cuarto y último lugar, y «digno de mención especial es el proyecto OPUS, probablemente la mayor colección de corpus paralelos multilingües de acceso libre.» (Doval, 2017) Su autor, Tiedemann, expone que la intención es abarcar docenas de lenguas e incluir múltiples áreas como derecho, administración, contenido audiovisual, software, artículos periodísticos, documentos médicos,etc. (2012) «[It] covers nearly 4,000 language pairs, for a total of over 40

billion words, distributed in approximately 3 billion parallel translation units (aligned sentences and fragments), and the collection is constantly growing.» (Gaspari, 2015)

Se debe tener en cuenta, no obstante, que los corpus mencionados no ofrecen una interfaz web para poder ser consultados en línea, por lo que muchos de ellos han de ser descargados en XML, entre otros formatos. Por último, cabe recalcar la imperiosa necesidad de que los materiales que conformen la base empírica de un trabajo de investigación lingüística cumplan con una alta calidad para asegurar los estándares establecidos. (Doval, 2017)

De igual importancia a todo lo anteriormente mencionado es el tratamiento de los textos, los cuales pasan por varios pasos antes de establecerse como corpus. Un primer paso es su selección y digitalización para, posteriormente, ser tratados. Una vez han sido seleccionados y digitalizados, estos se han de procesar de manera manual para prepararlos para el siguiente paso, la alineación. Este procesamiento, según detalla Doval (2017), «consiste básicamente en lograr tanto paralelismo como sea posible entre el texto fuente y el texto meta a fin de obtener los mejores resultados en la alineación automática», y por tanto, se eliminan todo lo que no sea texto y este se revisa. Al terminar su corrección, los textos se guardan en texto plano para su uso en la alineación. (Doval, 2017)

En lo que a la alineación se refiere, Tiedemann la define de la siguiente manera: «as a process of making symmetric correspondences explicit in order to enable further processing of parallel resources» (2011). Estas correspondencias se pueden establecer a diferentes niveles, desde palabras, hasta oraciones o párrafos completos. En la actualidad, la alineación estándar es a nivel de oración y, aunque aparentemente sencillo, este proceso es bastante complejo en la realidad. «Durante el proceso de traducción las oraciones pueden ser divididas, fusionadas, suprimidas, insertadas o reordenadas por el traductor para crear un texto natural» (Doval, 2017), lo que puede llegar a suponer todo un reto para la labor de alineación. No obstante, para solventar este dilema de manera automática, se han desarrollado alineadores automáticos.

6.2 Datos disponibles para la combinación es-nl

Pese a sus 25 millones de hablantes (Vermeiren, 2016), el neerlandés es considerado una lengua de menor difusión. Como consecuencia, uno de los problemas que afronta este idioma es «la poca rentabilidad económica de sus diccionarios u otras herramientas lingüísticas, tanto generales como terminológicos, monolingües o bilingües» (Vanden Bulcke y De Groote, 2016)

En el capítulo 1 se explicó la situación lingüística de los Países Bajos y de Bélgica, de forma que quedó clara la presencia predominante de lenguas como el francés, el alemán y el inglés debido

a su mayor peso demográfico y a su posición central en el continente. Por el contrario, lenguas menos centrales como el español o el italiano reciben menos atención, tanto a nivel académico como profesional. (Vermeiren, 2016)

Por otro lado, «la afinidad entre dos lenguas romances, como el francés y el español, es mayor que la que existe entre el español y una lengua germánica como el neerlandés» (Vermeiren, 2016), de manera que, compartiendo esta última oficialidad en el territorio junto a la lengua francesa, no es de extrañar que se cuenten con más datos en la combinación español-francés que español-neerlandés.

Una prueba de la falta de información y que, por tanto, apoya la motivación de este proyecto la presenta el departamento de Español de la Universidad de Nijmegen, quienes ya en 1999 tuvieron la iniciativa de crear un motor TA es-nl de carácter general. Sin embargo, el equipo declaró que la mayoría de los datos que se encuentran en tal combinación lingüística son de carácter técnico. (Hallebeek, 1999) Y es que los expertos coinciden en la necesidad de material suficiente para tales fines, y se recomienda que los datos recogidos no se limiten a dominios muy específicos para evitar complicaciones léxicas. (Hallebeek, 1999; Doval, 2017)

Respecto a la situación de la actividad dentro de las tecnologías de la traducción, la lengua neerlandesa ha estado incluida entre los idiomas de trabajo en proyectos académicos y de investigación. Concretamente, en lo que a sistemas de TA respecta, el neerlandés se combina con sus vecinos geográficos e históricos. Esto es, «English (39 systems), French (26 systems), and German (12 systems). Other frequent pairings are with Spanish (11), Italian (9), Russian (9), and Chinese (9) ». (Van der Beek y Van den Bosch, 2015)

Además, al ser el neerlandés una de las lenguas oficiales de la UE, está presente en las colecciones de Opus y de JRC-Aquis. (Van der Beek y Van den Bosch, 2015) Desafortunadamente, puesto que grandes cantidades de corpus español-neerlandés no existen y las existentes están muy especializadas, «constituía una necesidad la recopilación de un corpus paralelo alineado de tamaño y variedad léxica suficiente que nos suministrara la base empírica.» (Doval, 2017)

III. Marco metodológico

A continuación, se procede a explicar detalladamente cada uno de los pasos seguidos, desde la construcción desde cero de un corpus paralelo, hasta el entrenamiento de un motor TAN con los diferentes datos que se consiguen recabar. Finalmente, se realizará una evaluación del funcionamiento de cada entrenamiento y se hará una comparación de los datos.

1. Creación de un corpus paralelo

La creación de un corpus paralelo bilingüe debe seguir una serie de pasos para conseguir un material correcto, fiable y de calidad. Como se adelantó en apartados anteriores, el fin de este proyecto es suplir un vacío de datos en la combinación lingüística español-neerlandés, así como facilitar su aplicación en múltiples campos de investigación. Dichos pasos son los siguientes: (Martínez, 2003)

a. Selección de los textos
b. Anotación y etiquetación de los textos
c. Alineación de los textos

Se ha de tener en consideración, sin embargo, que constituye una compleja tarea que requiere de mucho tiempo y esfuerzo, así como de capacidad y medios informáticos para llevarla a cabo. (Doval, 2017)

1.2 Obtención de títulos paralelos

Para el presente proyecto se ha optado por obtener el corpus con el que se va a trabajar a partir de los artículos existentes en el portal de Wikipedia en las dos lenguas de trabajo.

Para ello, el primer paso era saber qué artículos de la Wikipedia en español corresponden con aquellos encontrados en la Wikipedia en neerlandés. Se usa entonces la plataforma de desarrollo colaborativo GitHub de donde se obtiene un código abierto disponible «Wikipedia Parallel Titles»[1] que se va a emplear para tal función.

[1] https://github.com/clab/wikipedia-parallel-titles

Este código utiliza dos archivos de los *dumps* de la Wikipedia[2]. Estos *dumps* son una recopilación de todos los datos y metadatos de la Wikipedia en todas las lenguas que se va actualizando con frecuencia.

Para utilizar el código de «Wikipedia Parallel Titles» se necesitan dos archivos por lengua de trabajo. Estos son los siguientes:

- «*nlwiki-20220301-page.sql.gz*», el cual contiene la información de los artículos de la Wikipedia en neerlandés junto con sus IDs.
- «*nlwiki-20220301-langlinks.sql.gz*», el cual contiene la información que permite conectar los artículos entre idiomas.

Se descargarán también, de esta forma, el par de archivos correspondientes de la Wikipedia en español. Con todos los archivos preparados, se pasa a ejecutar el código en la terminal de Ubuntu siguiendo las instrucciones indicadas en la plataforma.

Al finalizar el proceso, se obtiene un archivo al que se le nombra «es_nl_titles.txt» y que cuenta con un total de 593.356 títulos de artículos paralelos es-nl como se muestra a continuación:

```
Ruisvoorn ||| Scardinius erythrophthalmus
Ruit ||| Rombo
Ruit ||| Thalictrum
Ruiten Boer ||| Sota de Diamantes
Ruitenwisser ||| Limpiaparabrisas
Ruiter naar nationaliteit ||| Categoría:Jinetes por país
Ruiter van Artemision ||| Jinete de Artemisión
Ruiter van Madara ||| Caballero de Madara
Ruiter ||| Categoría:Jinetes
Ruiterstandbeeld koningin Wilhelmina ||| Estatua ecuestre de la reina Guillermina
Ruiterstandbeeld van Alexander III ||| Huevo de Alejandro III a caballo
Ruiterstandbeeld van Domitianus ||| Estatua ecuestre de Domiciano
Ruiterstandbeeld van Dzjengis Khan ||| Estatua ecuestre de Gengis Kan
Ruiterstandbeeld van Marcus Aurelius ||| Estatua ecuestre de Marco Aurelio
Ruiterstandbeeld van Trajanus ||| Equus Traiani
Ruiterstandbeeld van maarschalk Mannerheim ||| Estatua ecuestre del Mariscal Mannerheim
Ruiterstandbeeld ||| Categoría:Estatuas ecuestres
Ruiterstandbeeld ||| Escultura ecuestre
Ruitkrokodil ||| Crocodylus rhombifer
Ruitpython ||| Morelia spilota
Ruitpythons ||| Morelia
Ruitvlek-smalboktor ||| Stenurella bifasciata
Ruitvlekzalm ||| Hemigrammus caudovittatus
Ruitz ||| Ruitz
Ruitzalmen ||| Categoría:Citharinidae
Ruitzalmen ||| Citharinidae
Ruivos ||| Ruivos
Ruivães ||| Ruivães
Ruivós ||| Ruivós
Ruiz de Montoya ||| Ruiz de Montoya
Ruiz ||| Ruiz
Ruizia ||| Ruizia
Rujm al-Hīrī ||| Rujm el-Hiri
Ruk ||| Sobreaceleración
Ruka Norimatsu ||| Ruka Norimatsu
Rukai ||| Idioma rukai
Rukaj ||| Rukaj
```

Figura 4: Archivo «es_nl_titles.txt»

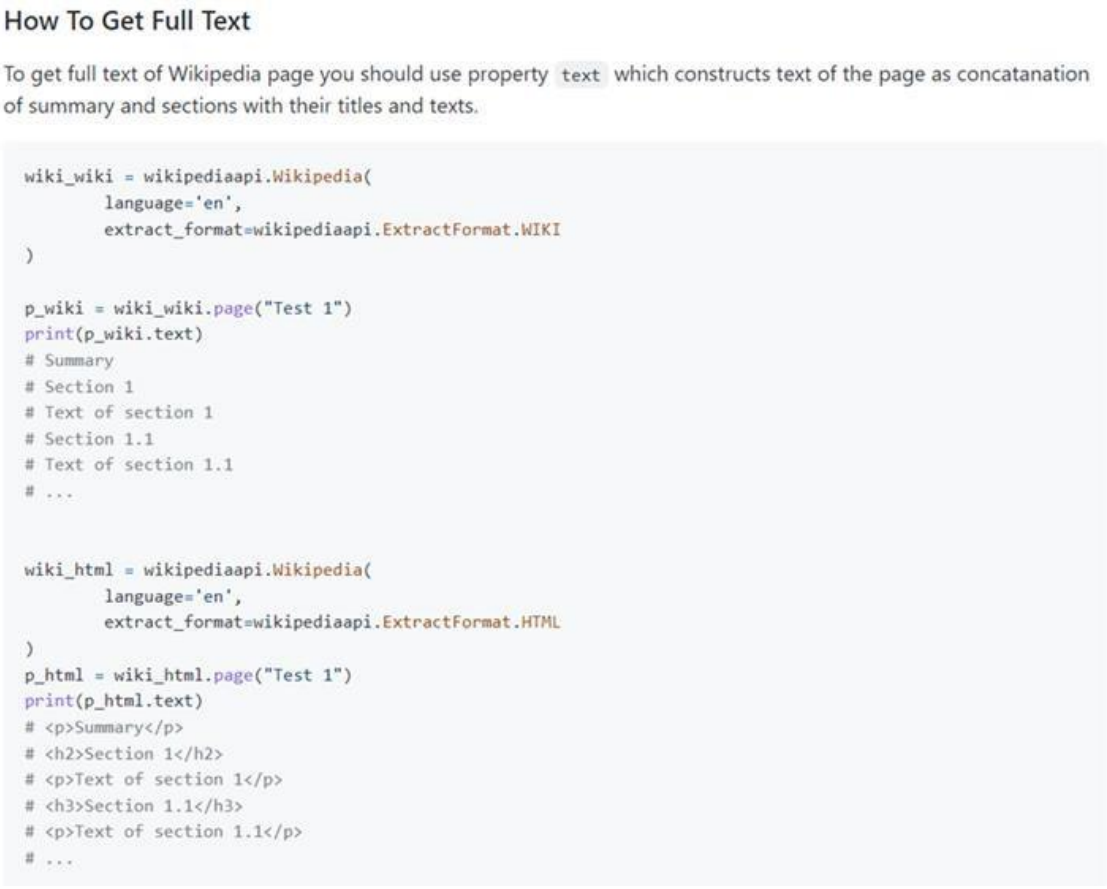
[2] https://dumps.wikimedia.org/

1.3 Obtención de artículos monolingües

Una vez recogidos todos los títulos paralelos, se necesita extraer el texto de cada archivo. Para ello, utilizamos dos métodos: Wikipedia API y el uso de *dumps*.

1.3.1 Wikipedia API

Se vuelve a utilizar un código ya escrito en GitHub[3] y, como lo que se necesita es conseguir todo el texto, se siguen los pasos del apartado «How To Get Full Text» que se muestra a continuación:



How To Get Full Text

To get full text of Wikipedia page you should use property `text` which constructs text of the page as concatanation of summary and sections with their titles and texts.

```
wiki_wiki = wikipediaapi.Wikipedia(
        language='en',
        extract_format=wikipediaapi.ExtractFormat.WIKI
)

p_wiki = wiki_wiki.page("Test 1")
print(p_wiki.text)
# Summary
# Section 1
# Text of section 1
# Section 1.1
# Text of section 1.1
# ...


wiki_html = wikipediaapi.Wikipedia(
        language='en',
        extract_format=wikipediaapi.ExtractFormat.HTML
)
p_html = wiki_html.page("Test 1")
print(p_html.text)
# <p>Summary</p>
# <h2>Section 1</h2>
# <p>Text of section 1</p>
# <h3>Section 1.1</h3>
# <p>Text of section 1.1</p>
# ...
```

Figura 5: Código de GitHub «How To Get Full Text»

Con este código, simplemente al introducir el título se obtiene todo el texto. Sin embargo, se necesitaría hacer el mismo proceso para obtener todos los títulos, algo que sería contraproducente y ralentizará todo el trabajo. Por tanto, se decide escribir un *script*.

Un *script* es un fragmento de código que se usa para hacer una repetición de acciones las veces que lo necesitemos. En este caso, lo que se ordena es sacar el texto de cada uno de los títulos

[3] https://github.com/martin-majlis/Wikipedia-API

sólo cuando haya contenido, tanto en los archivos de la Wikipedia en español como en los de la Wikipedia en neerlandés. Este paso es de suma importancia ya que no se cuenta con el mismo número de artículos en cada lengua y se correría el riesgo de obtener archivos vacíos. Por otro lado, se dan las indicaciones pertinentes para crear una carpeta de destino donde se guardarán los archivos producidos, y se indica el nombre y formato en que estos se han de guardar ('_nl.txt/'_es.txt).

Con el *script* creado en el procesador de texto NotePad++ y, por tanto, en formato .txt, se hace una conversión a formato *.py* y se ejecuta en la terminal de Ubuntu. El script utilizado recibe el nombre de «obtener_art_wiki.py» el siguiente:

```
#obrim l'arxiu amb els títols, fem dos variables: es_titles, nl_titles
#Per cada parella de títols, troba el text
# Obra dos arxius (es i nl) i escriu el text

from sentence_splitter import SentenceSplitter, split_text_into_sentences

import wikipediaapi

def get_text(language,title):
        wiki_wiki = wikipediaapi.Wikipedia(language=language,extract_format=
        wikipediaapi.ExtractFormat.WIKI)
        p_wiki = wiki_wiki.page(title)
        return(p_wiki.text)

titles = open('es_nl_titles.txt','r').read().splitlines()
titles_split = [title.split(' ||| ') for title in titles]

index = 1
for nl_title, es_title in titles_split:
    nl_text = get_text('nl', nl_title)
    es_text = get_text('es', es_title)
    if es_text != '' and nl_text != '':
        nl_file = open('output/'+str(index)+'_nl.txt','w')
        nl_sents = split_text_into_sentences(text=nl_text, language='nl')
        for nl_sent in nl_sents:
            if nl_sent != '':
                nl_file.write(nl_sent+'\n')
                es_file = open('output/'+str(index)+'_es.txt','w')
                es_sents = split_text_into_sentences(text=es_text, language='es')
        for es_sent in es_sents:
            if es_sent != '':
                es_file.write(es_sent+'\n')
        index = index+1
```

Figura 6: Archivo «obtener_art_wiki.py»

Problema principal durante la creación del corpus

Aparte de insignificantes inconvenientes mientras se generaba el *script*, se produce un verdadero problema durante este proceso. Y es que, a la hora de querer descargar todos los archivos necesarios, la página de la Wikipedia bloqueaba la dirección IP porque se estaban haciendo demasiadas solicitudes de descarga de las páginas a la vez.

Por otro lado, se calcularon 90 horas de carga con el *script* que, teniendo en cuenta el problema anterior, no iba a ser viable ya que se bloquearía el proceso en algún momento. Por tanto, se debía de buscar una vía diferente de actuación para conseguir los archivos, ya que solo se alcanzó descargar 8.000 archivos.

1.3.2 Extracción de texto de los dumps

Por consiguiente, se consideró que sería interesante la idea de crear el corpus a partir de los *dumps* de la Wikipedia. Esto permitiría descargar todos los archivos en local y así poder trabajar sin tener que depender de Internet.

Por tanto, siguiendo el mismo método de antes, se da con una forma de sacar los archivos de los *dumps*. Esto es, guardando el texto de los artículos de la Wikipedia a partir de su título, sólo si había contenido en ambas lenguas.

Se procede de la siguiente manera:

a. Primero, se obtienen los *dumps* que contienen el texto de los artículos:
   «eswiki-20220601-pages-articles.xml.bz2»
   «nlwiki-20220601-pages-articles.xml.bz2»
b. Luego, estos se convierten a formato JSON de manera que se pudiesen leer con el lenguaje de programación de Python. En este caso, la librería Gensim fue de gran ayuda siguiendo las instrucciones detalladas en el enlace[4].
c. Por último, se redacta de nuevo un *script* para obtener el texto deseado con las especificaciones necesarias para el trabajo.

Finalmente se obtienen un total de 321.971 títulos paralelos, 643.942 documentos en total.

1.4 Alineación de frases

Con el corpus de títulos paralelos preparado, un paso crucial en el proceso es la alineación. Se va a seguir el nivel de segmentación estándar hoy en día que es la alineación oracional, y para ello existen los llamados *aligners,* que son sistemas de alineación automática.

En un primer momento se decide utilizar Vecalign[5], pero se tuvo que descartar por limitación de recursos y por la lentitud del programa al tener que trabajar con tanta información.

[4] https://radimrehurek.com/gensim/scripts/segment_wiki.html
[5] https://github.com/thompsonb/vecalign

Tras otra búsqueda, se procede entonces a utilizar el alineador de BERTalign[6] en el entorno de Google Colab[7]. Este parecía tener un funcionamiento fácil, que trabaja con poco código, y solo había que sustituir las variables por los propios archivos de trabajo. Sin embargo, a pesar de que Google ofrece GPUs se llegó al límite de espacio. Se intentó subir los archivos en una carpeta comprimida *zip* pero el *script* no funcionaba correctamente.

Así, ante la ausencia de resultados, se optó por el uso de Hunalign, el cual se descartó desde el principio porque no contaba con un diccionario con la combinación lingüística es-nl, parte imprescindible en el proceso. Pese a ello, el elevado número de inconvenientes encontrados por el camino obligó a trabajar a contrarreloj y se decidió crear un diccionario sintético.

Un diccionario sintético es un diccionario falso creado a partir de otros diccionarios y, en este caso, se decidió utilizar el diccionario es-de, al ser el alemán una lengua germánica que, por tanto, se puede acercar al neerlandés.

Para ello, se crea un documento *.xlsx* con el contenido del diccionario y se realiza entonces la traducción del alemán al neerlandés con TA, de manera que quedan tres columnas como en la imagen siguiente:

D2 | = B2&" @ "&C2

	A	B	C	D	E
1	DE	NL	ES		
2	Ach	Ach	Vaya	Ach @ Vaya	
3	hallo	hallo	Hola	hallo @ Hola	
4	willkommen	welkom	bienvenidos	welkom @ bienvenidos	
5	ein	een	a	een @ a	
6	kein	geen	no	geen @ no	
7	der	de	la	de @ la	
8	dieser	deze	este	deze @ este	
9	jener	die	que	die @ que	
10	manche	sommige	alguno	sommige @ alguno	
11	genügend	genoeg	suficiente	genoeg @ suficiente	
12	viele	vele	muchos	vele @ muchos	
13	mehrere	meerdere	varios	meerdere @ varios	
14	verschiedene	verschillende	varios	verschillende @ varios	
15	andere	ander	otro	ander @ otro	
16	solche	zulke	tal	zulke @ tal	
17	jede	elk	cada	elk @ cada	
18	alle	allemaal	Todo el mundo	allemaal @ Todo el mundo	
19	beide	beide	ambas cosas	beide @ ambas cosas	
20	mein	mijn	mi	mijn @ mi	

Figura 7: Archivo .xlsx de diccionario sintético

[6] https://github.com/bfsujason/bertalign
[7] https://colab.research.google.com/?hl=es

Hunalign usa el formato *palabra en neerlandés @ palabra en español* para leer el contenido del diccionario, por lo que se añade la información de manera legible por el sistema.

Finalmente, el archivo resultante se convierte a formato de diccionario (esto es, *.dic*). Con esto ya se puede ejecutar el código de Hunalign sacado de GitHub[8] en la terminal Ubuntu según las instrucciones para proceder entonces con la alineación.

Sin embargo, aparece un pequeño inconveniente. Durante este procedimiento hay que volcar todos los archivos obtenidos en el paso anterior (esto quiere decir los 643.942 documentos en total) a un único archivo, ya que, al ser una gran cantidad de archivos, el terminal no permite ejecutarlo:

```
feedooms@DESKTOP-BQA3FES:/mnt/c/TFM/HUNALIGN/hunaligned/hunaligned$ cat * > ..alineacion-bruta.txt
-bash: /usr/bin/cat: Argument list too long
```

Figura 8: Error al cargar archivo

Finalmente, con el siguiente comando se soluciona el conflicto:

```
feedooms@DESKTOP-BQA3FES:/mnt/c/TFM/HUNALIGN/hunaligned/hunaligned$ find . -type f -name "*.txt" -exec cat {} + > ..alineacion_bruta.txt
cat: ./..alineacion_bruta.txt: input file is output file
```

Figura 9: solución error con doc de Hunalign

Tras ejecutar el código entero de Hunalign, se consigue un documento con la alineación de los textos en «sucio», archivos que deben ser tratados, como bien se comentó al principio de esta sección y gran parte del marco teórico, para poder servir de utilidad. Con este fin se deciden realizar dos pasos, uno para clasificar los archivos por calidad y otro que realizase una limpieza por idioma:

a. Filtro por confianza

Como se puede ver a continuación en un ejemplo en el que se expone el contenido exactamente como Hunalign lo genera, encontramos en primer lugar la frase en neerlandés; seguidamente, la oración en español; y, al final, el programa genera un número. Se muestran tres ejemplos:

1) Op 22 maart 2003 rijdt hij tijdens de kwalificatieritten voor de Grote Prijs van Maleisië de snelste tijd. El 22 de marzo de 2003 se convirtió en el piloto más joven en lograr

[8] https://github.com/danielvarga/hunalign

una ''pole position'' y un podio en Fórmula 1, con 21 años, en el Gran Premio de Malasia, pese a que aquel fin de semana estuvo con 39º de fiebre. 0.648062

2) Hij is onder meer bekend als dokter Carson Beckett in de televisieserie ''Stargate Atlantis''. Es más conocido por su papel como el Dr. Carson Beckett en la serie de televisión ''Stargate Atlantis''. 1.27073
3) Biografie Primeros años 0.18

Este número final es la confianza. Este valor indica la calidad de la alineación, es decir, en qué medida son equivalentes las frases alineadas.

Teniendo en cuenta dicho parámetro, Hunalign permite seleccionar el valor de confianza que tendrá como mínimo el archivo resultante, por lo que para el presente trabajo se decide elaborar dos versiones alineadas: una en la que el valor de confianza es 0, y otra en la que sea 0.85. De esta forma, se puede entrenar un motor y comparar los resultados, para observar si varían mucho estos, dependiendo de qué confianza se le aplica.

Como resultado de este proceso, se obtienen dos corpus: corpus de confianza 0, y corpus de confianza 0,85. Por otro lado, se seleccionan los 100.000 archivos con mayor puntuación para que constituya otro corpus independiente.

a. Filtro por idioma

Al revisar los documentos, se observó que estos contaban con muchos segmentos que, o bien no tenían contenido lingüístico, o bien estaban en inglés, por lo que se toma la decisión de hacer un filtrado por idioma para limpiarlos. Esto se consigue con el siguiente *script* «filer_by_lang.py»:

```
import langid

# Open file with alignments and scores
file = open('hunaligned.txt','r').read().splitlines()

# Open files to write
file_out_nl = open('hunaligned_lang_filter.nl','w')
file_out_es = open('hunaligned_lang_filter.es','w')
file_out_scores = open('hunaligned_lang_filter.scores','w')

# Process all lines and append them to a list
count = 0
all_lines = []
for line in file:
    if len(line.split('\t')) == 3: # To avoid parsing errors
        nl, es, score = line.split('\t')
        nl_lang = langid.classify(nl)[0]
        es_lang = langid.classify(es)[0]
        if nl_lang == 'nl' and es_lang == 'es':
            if len(nl.split())>=10 and len(es.split()) >= 10: # Sentences with longer than 10 words
                file_out_nl.write(nl+'\n')
                file_out_es.write(es+'\n')
                file_out_scores.write(score+'\n')
    if count % 100 == 0:
        print('{} sentences processed.'.format(count))
    count += 1
```

Figura 10: Archivo «filer_by_lang.py»

Por otro lado, se eliminan los fragmentos repetidos ya que no interesa trabajar con información duplicada. Finalmente, y de cara a poder entrenar el motor, se deben separar los archivos alineados obtenidos en dos documentos diferentes, uno para cada lengua, ya que constituye un requisito indispensable en el entrenamiento.

Como resultado, se obtiene otro corpus a partir del corpus con los 100 mejores resultados, pero esta vez filtrado por idioma. Este recibe el nombre de «100k highest scores lang», y pondría fin a las colecciones de trabajo con las que se procederá a entrenar el motor.

Una vez conseguido esto, los corpus creados y separados por idioma se deben subir a la plataforma que veremos a continuación, siendo imprescindibles para la correcta ejecución del programa.

2. Entrenamiento de un motor

Llegado este punto, se cuenta con cuatro corpus de diferentes características con los que realizar el entrenamiento: el corpus con los 100.000 archivos con mayor puntuación en Hunalign, este mismo, pero filtrado por idiomas, aquel corpus que ha obtenido confianza 0, y el que obtuvo confianza 0,85. De esta manera, se pasa a trabajar directamente con la plataforma MutNMT para hacer las pruebas.

2.1 La plataforma MutNMT

Como se expone previamente en la parte teórica del presente trabajo, el motor que se va a entrenar es el MutNMT.

MutNMT es una aplicación web para entrenar motores neurales de motores de traducción automática con fines didácticos. Actualmente los profesores y estudiantes interesados en utilizar la herramienta pueden acceder a MutNMT a través de la interfaz de la UAB[9], como por la UGA[10]. El acceso se realiza con cualquier cuenta de Gmail. Además, el código de la aplicación está disponible en GitHub[11] de nuevo.

MutNMT permite al usuario entrenar, inspeccionar, evaluar y traducir con motores TAN.

[9] https://ntradumatica.uab.cat
[10] http://multitrainmt.univ-grenoblealpes.fr:5000
[11] https://github.com/Prompsit/mutnmt

Esta es la interfaz principal de la plataforma de MutNMT:

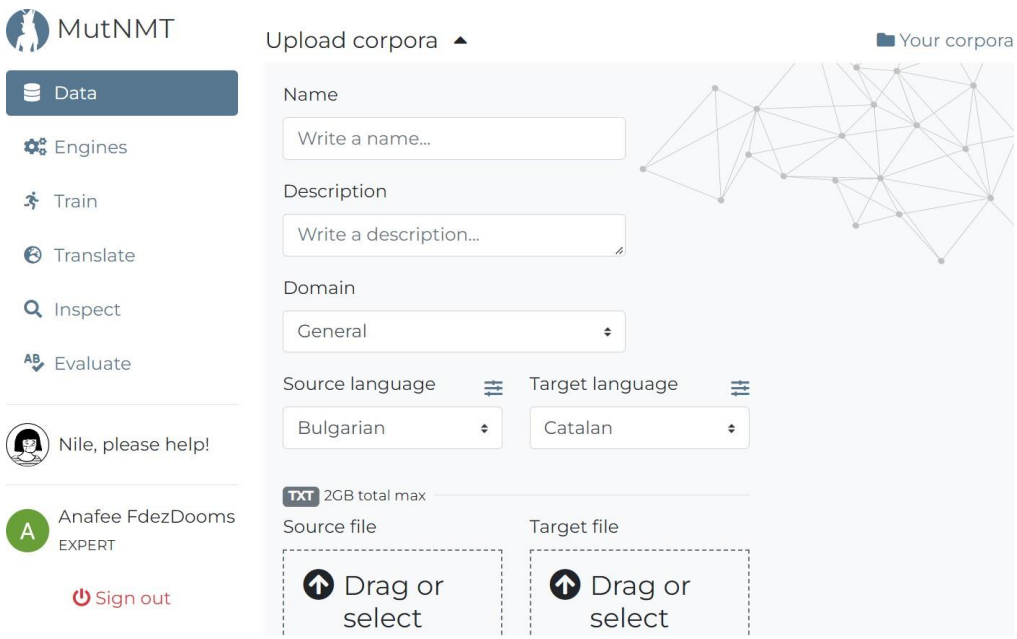


Figura 11: Interfaz de la plataforma MutNMT

En la primera pestaña de «Data» es donde se suben los datos de los corpus creados. Esto se puede hacer tanto en formato *.txt* como en TMX. En caso de hacerlo en formato texto, se sube el archivo en ambos idiomas por separado.

Finalmente se ha optado por subir el corpus paralelo con confianza 0,85 y el que tiene la puntuación más alta, esto es, los 100.000 archivos con la puntuación más alta.

Para poder hacer una comparación con los datos obtenidos de los corpus nuevos, se ha entrenado un motor a base de un corpus multilingüe obtenido automáticamente de la Wikipedia, llamado Wikimatrix y que está disponible en el repositorio de corpus OPUS[12]. Se hace el entrenamiento igual con los datos que se pueden descargar públicamente, aunque en este caso será en formato TMX. Esto solo cambia el número de archivos que haya que subir, por lo que hay que subir un solo archivo, en lugar de uno para cada idioma.

Una vez subidos los archivos en «Data», se puede empezar a entrenar el motor. Esto se hace desde la pestaña «Train»:

[12] https://opus.nlpl.eu/

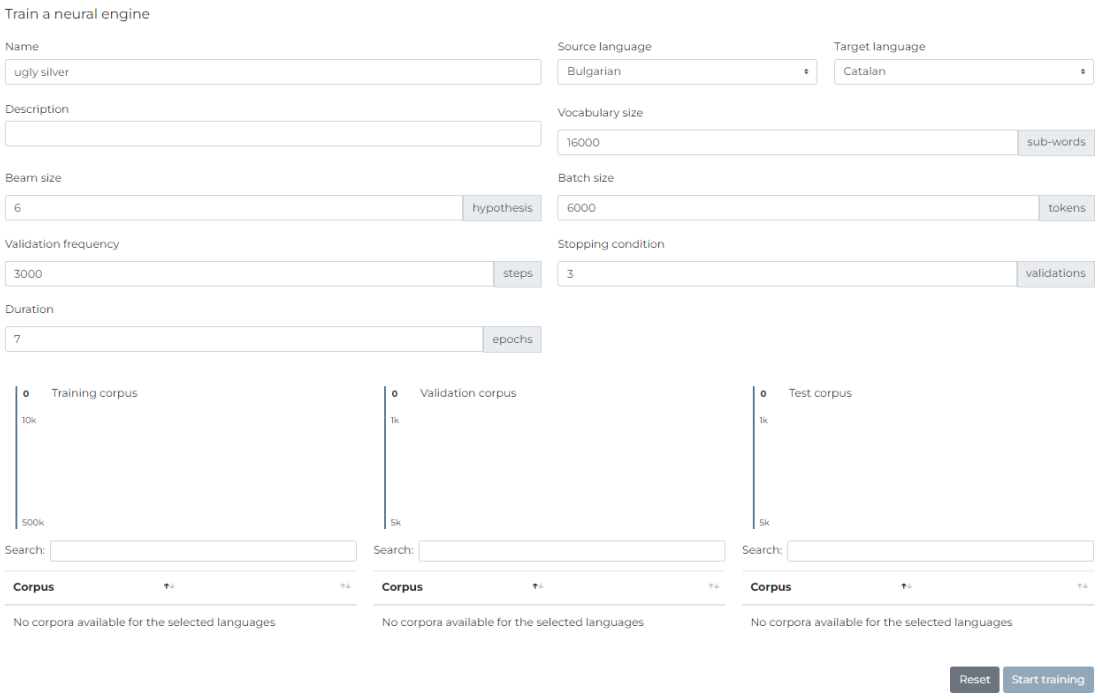


Figura 12: Pestaña de «Train» de MutNMT

En la imagen que se muestra a continuación aparece el apartado desde donde se pueden elegir los datos y los corpus con la que se va a trabajar, entre otros detalles:

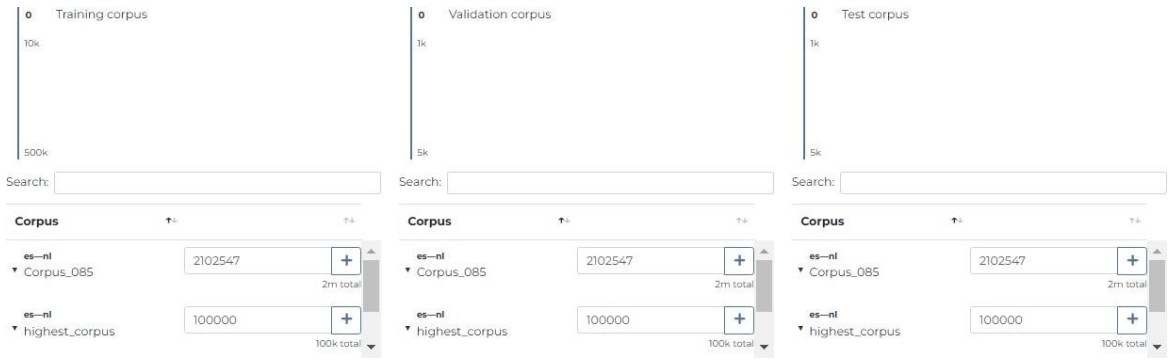


Figura 13: Selección de los corpus en MutNMT

Cuando se introduce la combinación lingüística que se quiere entrenar, salen todos los corpus que se han subido a la página, y es cuestión de elegir la cantidad de cada corpus para entrenar el motor.

A pesar de que existe la opción de escoger algunos parámetros, el objetivo del trabajo es comparar la calidad de los corpus recogidos, así que se mantienen los parámetros por defecto para entrenar modelos que sean consistentes y comparables al mismo nivel.

2.2 Motores entrenados

Como se expone anteriormente, se han entrenado cuatro motores diferentes: a partir del corpus de confianza 0.85, de confianza 0, con el corpus de los 100.000 con mejor puntuación, y este último, pero con los idiomas filtrados.

Esta es la pestaña que ofrece la plataforma MutNMT cuando se finaliza un entrenamiento. De nuevo, aparecen todos los parámetros seleccionados para el entrenamiento en cuestión, así como los tiempos empleados y las características de los corpus elegidos con anterioridad. En este caso se expone con el corpus de los 100k con mejor puntuación a modo de ejemplo:

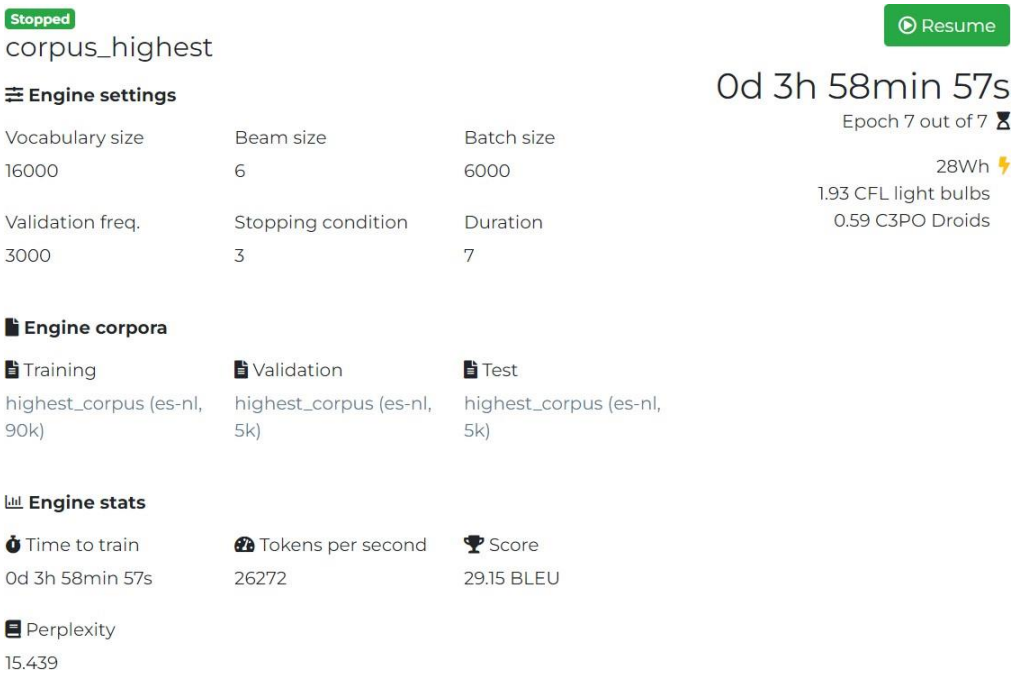


Figura 14: Pestaña de resultados en MutNMT

Aquí se puede ver toda la información sobre el entrenamiento del motor. Como se observa, pulsando en el botón de «Resume» se podría continuar con el entrenamiento. De por sí, el entrenamiento es siempre de una hora, y se puede prolongar en el tiempo para que vaya

mejorando. Cada vez tardará menos y finalmente llegará un punto en que no se permita seguir entrenándolo. Una vez haya finalizado, se pondría a prueba los resultados a través de una traducción. Esto se hace en la pestaña «Translate». Ahí mismo se elige qué motor se quiere emplear para la tarea.

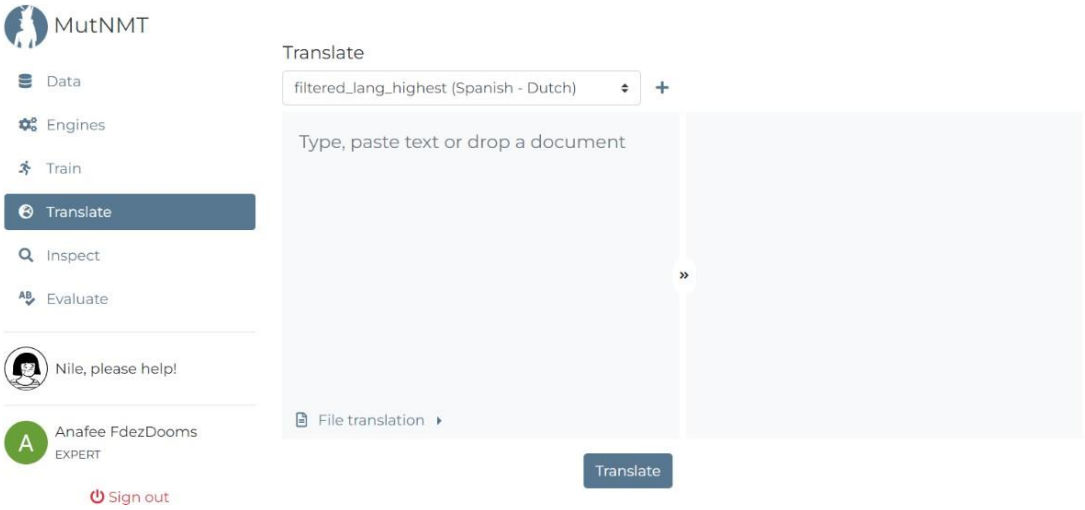


Figura 15: Pestaña de «Translate» en MutNMT

3. Evaluación

Finalmente, y tras subir el archivo que se quiere traducir, se eligen los parámetros para la evaluación del motor empleado:

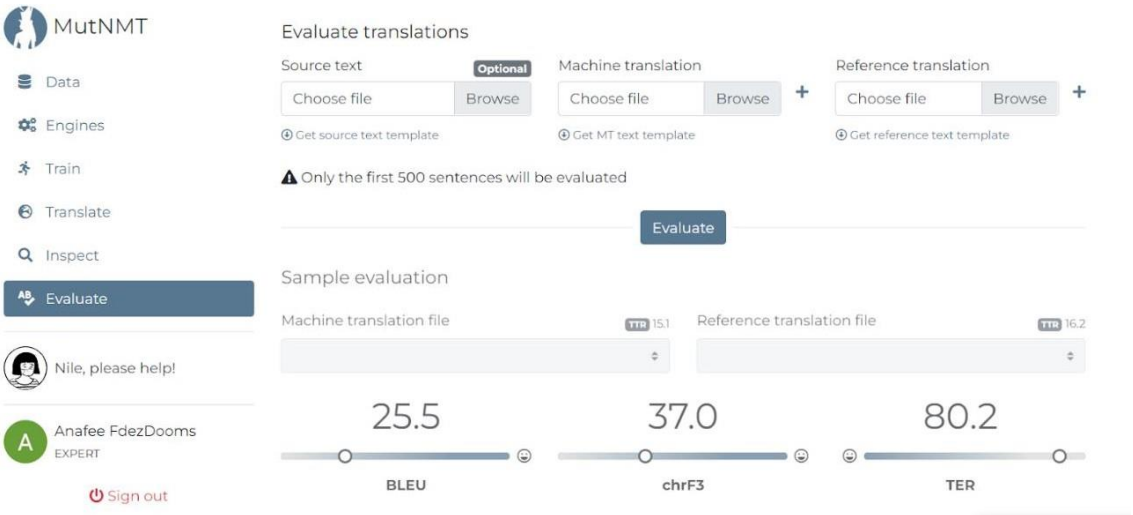


Figura 16: Pestaña de evaluación de MutNMT

Para evaluar nuestro motor, se necesita un *dataset* con el mismo archivo en los dos idiomas de trabajo.

Se usa el *dataset* de Flores-101[13], un *test set* de referencia desarrollado por Facebook de 3.001 oraciones extraídas de la Wikipedia en inglés que cubren una amplia variedad de temas. Estas frases han sido traducidas por traductores profesionales a 101 diferentes idiomas. Por esta razón, se usa como *dataset* de referencia y de evaluación.

Los archivos de Flores se usan tanto como texto original para obtener la primera traducción automática, como para luego evaluar esta misma.

Al haber usado el archivo de referencia en español del *dataset* de Flores, se necesita también el archivo traducido al neerlandés.

En la pestaña «Evaluate», se permite evaluar los modelos con un *test set* distinto. Se sube el archivo original en español de Flores, la traducción automática obtenida con el motor entrenado y el archivo traducido de Flores en neerlandés.

Cuando acaba de hacer la evaluación, que usa como referencia los archivos originales y traducidos que se suben, la plataforma proporciona una puntuación en la métrica BLEU, que se comentará en la siguiente sección.

[13] https://ai.facebook.com/tools/flores/

IV. Resultados

En este apartado se incorporarán todos los resultados obtenidos en los diferentes procesos que se han llevado a cabo a lo largo de todo el trabajo. Se proporcionará la información recogida en tablas para ofrecer una visión más clara de los análisis contrastivos.

En esta primera tabla, se recogen el número de documentos, frases y tokens que hay en los corpus comparables creados. El primero, «es_nl_titles.txt», es el que contiene los títulos paralelos sin texto de la Wikipedia. En el segundo, se encuentran los documentos resultantes de aplicar el *script* «obtener_art_wiki.py», con texto de la Wikipedia incluido. En estos, al ser 643.942 documentos diferentes y no solo uno, no se ha podido realizar un recuento detallado.

Corpus comparables	Documentos	Frases	Tokens
es_nl_titles.txt	593.356	613.057	3.284.845
obtener_art_wiki.py	321.971 en total, 643.942 títulos paralelos	-	-

Tabla 1: Corpus comparables

En la siguiente tabla, se recogen los datos de los cuatro corpus paralelos finalmente creados, incluyendo el número de segmentos, número de tokens en español, número de tokens en neerlandés, y el número de tokens encontrados por segmento.

Corpus paralelos	Segmentos	Tokens ES	Tokens NL	Tokens/Segmento
Corpus 100k highest scores lang	1.468.943	34.890.382	27.695.323	23,75
Corpus 100k highest scores	100.000	2.232.743	1.768.725	22,32743
Corpus confianza 0,85	2.102.547	23.346.040	21.229.080	11,10
Corpus confianza 0	7.881.638	94.922.473	91.433.314	12,04

Tabla 2: Corpus paralelos

Se puede observar que, en el corpus de confianza 0, hay cantidades considerablemente mayores que en los demás debido a que, con una puntuación tan baja, se da pie a mayor número de resultado. Por otro lado, se aprecia que el corpus con confianza 0,85 obtiene menos de la mitad que este último, por lo que debería ser mejor el motor entrenado a base de este corpus.

Con estos resultados se puede ver, además, que el corpus con la selección de los segmentos con mayor calidad filtra frases cortas, ya que el número de tokens por oración es mucho mayor.

En la tercera tabla que se presenta a continuación se encuentra la información detallada sobre los resultados obtenidos con el uso del alineador Hunalign.

Títulos obtenidos	613.057
Documentos comparables	643.942
Segmentos alineados	11.553.622
Segmentos alineados con confianza 0	7.881.638
Segmentos alineados con confianza 0.85	2.102.547
Segmentos alineados filtrados por idioma	1.468.943
100k highest scores	100.000
100k highest scores filtrados por idioma	520.000

Tabla 3: Datos Hunalign

Respecto a los resultados obtenidos en BLEU en la plataforma MutNMT, se recoge la información de la tabla 4. Como era de esperar, el corpus que obtiene la puntuación más alta es el que está compuesto por los 100.000 segmentos con mayor puntuación por Hunalign.

Por el contrario, cuando se evalúa con el *dataset* de Flores, se puede observar en la tabla 5 que baja considerablemente esta puntuación de manera general, habiendo pasado por el mismo entrenamiento y el mismo *dataset*.

Corpus 100k highest scores	29.15 BLEU
Corpus 0,85 confianza	12.77 BLEU
Corpus Wikimatrix	16,1 BLEU
Corpus 100k highest scores lang	5,39 BLEU

Tabla 4: Puntuaciones BLEU

Corpus 100k highest scores	3 BLEU
Corpus 0,85 confianza	0,5 BLEU
Corpus Wikimatrix	9 BLEU
Corpus 100k highest scores lang	2,9 BLEU

Tabla 5: Puntuaciones BLUE de Flores 101

En este punto cabe realizar ciertas aclaraciones:

Como se puede observar, ambas tablas incorporan información del corpus Wikimatrix que se decidió añadir para hacer una comparación con relación a la calidad. Y, a decir verdad, en cuanto a puntuación general, este no puntúa mucho más alto que el resto de los corpus. Sin embargo, cuando se trata de su uso para el *dataset* de Flores, es el corpus que tiene la puntuación más alta. Esto puede ser debido a que el corpus de la Wikimatrix contiene alrededor de 250 millones de tokens, lo que supera en gran medida los datos que se han conseguido recabar para la creación de los corpus propios. Asimismo, se ha decidido no contar con el corpus de confianza 0 por razones evidentes, ya que se dio por sentado que sus resultados no entrarían en un análisis contrastivo de calidad.

Por otro lado, se ha incorporado el corpus «100k highest scores lang», el cual es igual que el corpus «100k highest scores» pero filtrado por idiomas. Este ha sido de especial dificultad poderlo entrenar debido a problemas técnicos con la plataforma, pero finalmente se consiguieron los resultados. Para sorpresa del equipo, el corpus es el que ha sacado la peor puntuación de los cuatro en ámbito general, y se posiciona en tercer lugar al aplicarse al *dataset* de Flores. Si se tienen en cuenta sus características, debería de puntuar por encima del corpus «100k highest scores», ya que se elaboró con el objetivo de superar a este. Con esto, el equipo considera que debe de haber un fallo en los datos internos del corpus que entorpezca el buen funcionamiento a la hora de ser entrenado.

Por otro lado, la siguiente imagen corresponde a la puntuación que recibe el corpus de Wikimatrix por la traducción del *dataset* de Flores en la plataforma MutNMT:

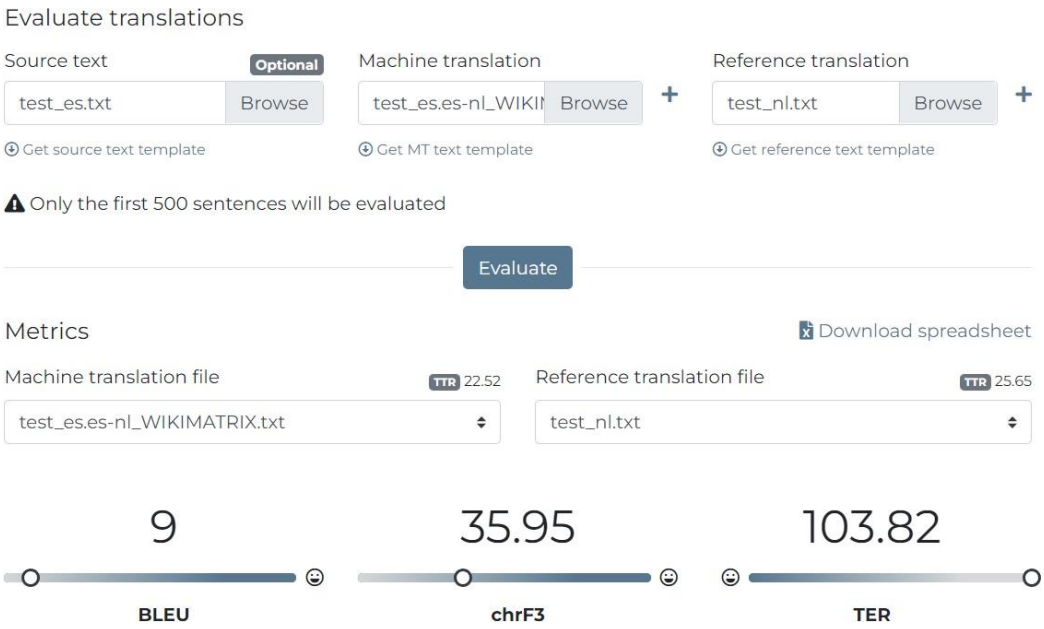


Figura 17: Resultados corpus Wikimatrix en MutNMT

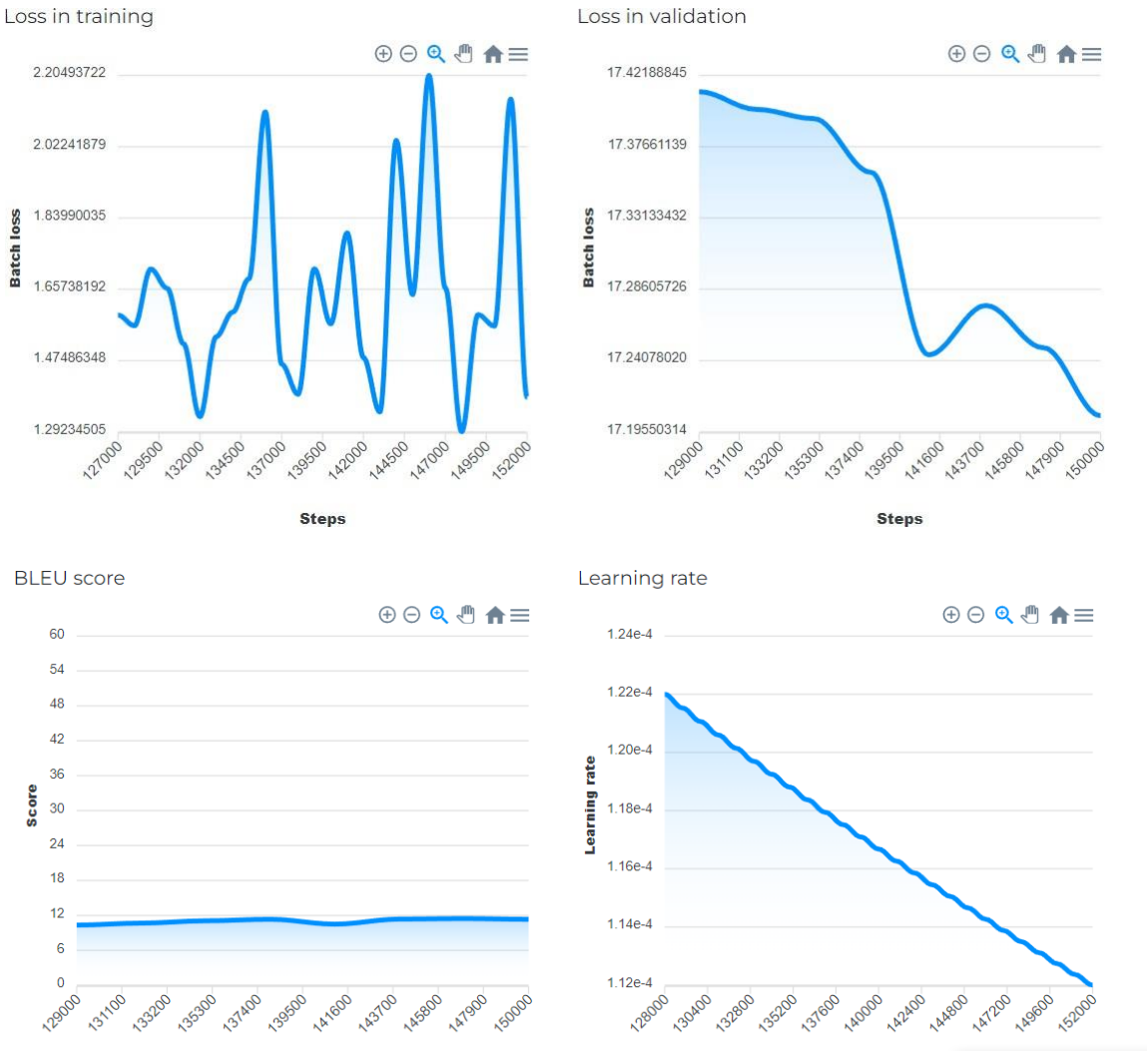
Por último, se incorporan seguidamente cuatro gráficas proporcionadas por la misma plataforma MutNMT con datos sobre el rendimiento del entrenamiento de cada uno de los corpus. Estas gráficas representan la pérdida de datos durante el entrenamiento, la pérdida en la valoración, la puntuación de BLEU y el índice de aprendizaje.

- Corpus 100k highest scores

Gráfica 1: Gráficas del corpus 100k highest scores

Se puede observar una gran bajada en el entrenamiento de este motor. Si se mira el resultado del siguiente motor, el que está basado en el corpus de confianza 0,85, se ve que no pasa lo mismo. Se podría pensar que esto es debido a una peor calidad de los datos usados.

- Corpus confianza 0,85



Gráfica 2: Gráficas del corpus confianza 0,85

- Corpus Wikimatrix

Gráfica 3: Gráficas del corpus Wikimatrix

- Corpus 100k highest scores lang

Gráfica 4: Gráficas del corpus 100k highest scores lang

Una vez argumentados con evidencia gráfica los resultados de cada uno de los motores que nos ocupaban en este trabajo, se estima oportuno comentar las limitaciones encontradas a lo largo del proceso. Por último, se responderán las preguntas de investigación que se plantearon en la primera sección.

En primer lugar, ya se atisbaron inconvenientes en la recogida de datos, para la cual el equipo era conocedor de la escasez de datos en la combinación elegida después de realizar un breve estudio sobre la situación lingüística. Se intentó recabar la información de mayor calidad que estuviese disponible para el público general, pero, tras ser evaluada por diferentes métodos, es

evidente que no se puede asegurar los estándares de calidad de los corpus empleados. Por tanto, se puede afirmar que el presente trabajo se planteaba como un reto desde el comienzo.

En segundo lugar, lo que ha supuesto un punto de inflexión en cuanto a la calidad del presente trabajo ha sido el proceso de alineación. El equipo no contaba con que iba a tener que dedicar un tiempo excesivamente largo a esta parte del proceso y que, además, no iba a alcanzar los resultados que se perseguían. Estos problemas se deben principalmente a razones computacionales, ya que no se contaba con acceso a ordenadores con interfaz gráfica lo suficiente potentes como para llevar a cabo tales procedimientos. Por otro lado, pese a la gran formación con la que el equipo cuenta y a la ayuda de profesionales, no se disponían de las habilidades técnicas suficientes para suplir de manera satisfactoria los problemas técnicos encontrados por el camino. De esta manera, y aun habiendo encontrado soluciones que permitiesen finalizar el proyecto, como por ejemplo la creación y uso de un diccionario sintético ante la falta de recursos, los índices de calidad no son los que se esperaban.

En tercer y último lugar, mencionar la funcionalidad de la plataforma MutNMT. Al ser un proyecto relativamente novedoso y con un enfoque didáctico, la plataforma cuenta con una interfaz muy intuitiva y fácil de usar, pero tiene sus limitaciones.

Antes de pasar a las preguntas de investigación, otro factor a tener en cuenta y que no se debe pasar por alto es la posición que ocupa la TAN en las investigaciones actuales, ya que, siendo un campo de investigación en pleno desarrollo, aún quedan terrenos por explorar en los que ya los grandes profesionales están volcados. Asimismo, mencionar que este estudio constituye el Trabajo Final de Máster, por lo que no se cuenta con fondos económicos, tan necesarios en la elaboración de un motor TAN como bien se expone en capítulos anteriores.

Finalmente, se contestan las preguntas de investigación en base a los conocimientos adquiridos, tanto a nivel de documentación como a nivel práctico:

1. ¿Es mejor un corpus con una gran cantidad de datos y de mala calidad o un corpus con menos datos, pero corregido y de calidad?

Realmente, se necesita el equilibro de ambos valores. Por un lado, se debe de contar con vastas cantidades de corpus, esto es, millones de segmentos alineados. Por el otro, estos datos deben ser de máxima calidad y deben de haber pasado por un control exhaustivo previo a ser puesto en marcha. En definitiva, en lo que respecta a los motores TAN, para conseguir un aprendizaje fiable, el motor se debe entrenar con la mayor cantidad de datos posibles y que cumplan con los estándares de calidad establecidos.

2. ¿Cuál es el mínimo de datos necesarios para poder obtener un corpus de calidad?

Como bien se expone en la pregunta anterior, la cual está directamente relacionada con esta, no existe un mínimo necesario para poder obtener un corpus de calidad. Ahora bien, mientras mayor sea el número de datos, mayor será el aprendizaje del motor. Si lo que se busca son resultados fiables, de nuevo, estos han de ser de exquisita calidad.

V. Conclusiones

Como recapitulación del trabajo en general y teniendo en cuenta los retos afrontados y las limitaciones encontradas, se obtienen en claro una serie de conclusiones.

Como bien se reitera en múltiples ocasiones y en base a los resultados de la parte empírica del trabajo, se desmiente la falsa asunción de que el traductor profesional será reemplazado por la traducción automática. No solo para la creación del motor, sino a lo largo de todo el proceso de traducción se necesita la supervisión de un lingüista especializado con conocimientos avanzados de informática que aseguren los estándares de calidad de las traducciones finales procesadas por motores de traducción automática.

En cuanto a los objetivos marcados, se demuestra la imperiosa necesidad de la creación de un corpus de las características que para este trabajo se necesitaban y se confirma la extrema escasez de datos en la combinación español-neerlandés con la que se cuenta hoy en día, hasta el punto de no existir siquiera un diccionario bilingüe de calidad fiable con la que entrenar un motor. Asimismo, se ha conseguido contribuir al campo de investigación con un corpus paralelo es-nl de calidad revisada y controlada del que se puede partir para futuros proyectos.

Por otro lado, no se han podido alcanzar los objetivos que perseguían la creación de un motor TAN eficiente ya que su entrenamiento dependía directamente de los datos que se incorporasen, y estos no llegaban al mínimo necesario para aportar resultados fiables. Por otro lado, se han dado ciertos problemas técnicos en la plataforma de desarrollo fuera de las competencias que al equipo les pertenece que han dificultado el entrenamiento del motor con los corpus propios. En general, se contaba con la premisa de que los resultados iban a ser altamente positivos, pero, con toda la problemática comentada anteriormente, se deberá posponer para futuros proyectos.

Aun con esto, el proyecto ha superado las expectativas de aprendizaje del equipo ya que, al tener que superar una cantidad de inconvenientes superior a la estimada, este ha tenido que trabajar y reforzar sus habilidades de programación para ofrecer soluciones aceptables. Se ha conocido, además, de primera mano como funcionan los motores TAN y lo que supone la creación de un corpus paralelo partiendo desde cero. Por último, cabe mencionar los conocimientos adquiridos a partir de magníficas lecturas de profesionales del círculo las cuales han sido de valiosa ayuda para poder llevar a cabo el trabajo.

En cuanto a futuras líneas de investigación, se proponen los siguientes puntos:

- Inducir diccionarios a partir de los corpus paralelos existentes
- Buscar métodos para conseguir una alineación más sofisticada y de calidad
- Entrenar un motor TAN con datos más fiables y de mejor calidad

Este trabajo busca incentivar y promover la investigación en la lingüística computacional también de lenguas minoritarias. En plena era de la globalización donde la tecnología es el brazo que acerca las diferentes culturas, no se debe dar la espalda a aquellas lenguas que no cuentan con los suficientes recursos para establecer una comunicación directa con el resto del mundo. El objetivo principal era hacer una aportación para facilitar esa comunicación entre las lenguas española y neerlandesa, por lo que por la presente se dan los primeros pasos en dicha dirección.

VI. Bibliografía

ANDÚJAR, Á. (2021). *TRADUCCIÓN AUTOMÁTICA NEURONAL SENSIBLE AL CONTEXTO*. TRABAJO INÉDITO DE FIN DE MÁSTER. UNIVERSIDAD POLITÉCNICA DE VALENCIA

ARNOLD, D., BALKAN, L., HUMPHREYS, R. L., MEIJER, S., Y SADLER, L. *MACHINE TRANSLATION: AN INTRODUCTORY GUIDE*.

BEEK, L. V. D., Y VAN DEN BOSCH, A. P. J. (2015). "TRANSLATION TECHNOLOGY IN THE NETHERLANDS AND BELGIUM". *ROUTLEDGE ENCYCLOPEDIA OF TRANSLATION TECHNOLOGY,* PP. 390-401. ROUTLEDGE

CASABUERTA, F. Y PERIS, A. (2017). "TRADUCCIÓN AUTOMÁTICA NEURONAL". REVISTA TRADUMÀTICA. TECNOLOGIES DE LA TRADUCCIÓ, 15, PP. 66-74.

CASTILHO, S., MOORKENS, J., GASPARI, F., CALIXTO, I., TINSLEY, J., & WAY, A. (2017). IS NEURAL MACHINE TRANSLATION THE NEW STATE OF THE ART? THE PRAGUE BULLETIN OF MATHEMATICAL LINGUISTICS, 108(1). [DISPONIBLE EN: https://doi.org/10.1515/pralin-2017-0013.]

CHÉRAGUI, M. A. (2012). "THEORETICAL OVERVIEW OF MACHINE TRANSLATION" *ICWIT*, PP. 160-169.

DECLERCQ, C. (2015). "EDITING IN TRANSLATION TECHNOLOGY". *ROUTLEDGE ENCYCLOPEDIA OF TRANSLATION TECHNOLOGY,* PP. 364-374. ROUTLEDGE

DENKOWSKI, M., Y LAVIE, A. (2011). "METEOR 1.3: AUTOMATIC METRIC FOR RELIABLE OPTIMIZATION AND EVALUATION OF MACHINE TRANSLATION SYSTEMS". *PROCEEDINGS OF THE SIXTH WORKSHOP ON STATISTICAL MACHINE TRANSLATION*, PP. 85-91.

DONALDSON, B. C. (1983). DUTCH: A LINGUISTIC HISTORY OF HOLLAND AND BELGIUM. M. NIJHOFF.

DOVAL, I.(2017). "LA CONSTRUCCIÓN DE UN CORPUS PARALELO BILINGÜE MULTIFUNCIONAL". MOENIA, 23.

FORCADA, M. L.; PÉREZ, L. & RICO, C. (2018) "TRANSLATORS' TRACK: A GLOSSARY". *ALICANTE: EAMT*. [DISPONIBLE EN: ELECTRÓNICA: <HTTP://EAMT.ORG/TRANSLATORS_DOCUMENTS/EAMT_2018_GLOSSARY.PDF>.

FORCADA, M. L. (2017). "MAKING SENSE OF NEURAL MACHINE TRANSLATION". *TRANSLATION SPACES*, 6(2), 291-309.

FORCADA, M. L. (2014). "OPEN-SOURCE MACHINE TRANSLATION TECHNOLOGY". *ROUTLEDGE ENCYCLOPEDIA OF TRANSLATION TECHNOLOGY*, PP. 190-204. ROUTLEDGE.

FORCADA, M. L. (2010). "MACHINE TRANSLATION TODAY". *HANDBOOK OF TRANSLATION STUDIES*, 1, 215-223.

GARCÍA, C. (2019). *CREACIÓN DE UNA GUÍA PRÁCTICA DE TRADUCCIÓN AUTOMÁTICA NEURONAL PARA ESTUDIANTES DE TRADUCCIÓN.* TRABAJO INÉDITO DE FIN DE GRADO. UNIVERSIDAD DE ALICANTE.

GARCÍA, I. (2015). "COMPUTER-AIDED TRANSLATION: SYSTEMS". *ROUTLEDGE ENCYCLOPEDIA OF TRANSLATION TECHNOLOGY*, PP.68-87. ROUTLEDGE.

GASPARI, F. (2014). "ONLINE TRANSLATION". *ROUTLEDGE ENCYCLOPEDIA OF TRANSLATION TECHNOLOGY,* PP. 616-631. ROUTLEDGE.

GINESTÍ, M., Y FORCADA, M. L.. (2009). "LA TRADUCCIÓ AUTOMÀTICA EN LA PRÀCTICA: APLICACIONS, DIFICULTATS I ESTRATÈGIES DE DESENVOLUPAMENT". *CAPLLETRA. REVISTA INTERNACIONAL DE FILOLOGIA,* (46), 43-60.

GÓMEZ LOSCOS, A., Y GONZÁLEZ SANZ, M. J. (2014). "LA EVOLUCIÓN RECIENTE DEL TURISMO NO RESIDENTE EN ESPAÑA". *BOLETÍN ECONÓMICO/BANCO DE ESPAÑA*, PP. 67-74.

HALLEBEEK, J. (1999). EL CORPUS PARALELO. PROCESAMIENTO DEL LENGUAJE NATURAL. UNIVERSIDAD DE NIJMEGEN

HUTCHINS, J. (2014). "THE HISTORY OF MACHINE TRANSLATION IN A NUTSHELL". *RETRIEVED DECEMBER*, 20(2009), 1-1.

HUTCHINS, J. (2010). "MACHINE TRANSLATION: A CONCISE HISTORY". *COMPUTER AIDED TRANSLATION: THEORY AND PRACTICE*, 13(PP.29-70), 11.

HUTCHINS, J. (2009). "MULTIPLE USES OF MACHINE TRANSLATION AND COMPUTERISED TRANSLATION TOOLS". *MACHINE TRANSLATION*, PP. 13-20.

HUTCHINS, J. (2005). "TOWARDS A DEFINITION OF EXAMPLE-BASED MACHINE TRANSLATION". *WORKSHOP ON EXAMPLE-BASED MACHINE TRANSLATION,* PP. 63-70.

LAN, L. (2014). "CORPUS". *ROUTLEDGE ENCYCLOPEDIA OF TRANSLATION TECHNOLOGY*, PP. 503-517. ROUTLEDGE.

LÓPEZ, A. (2018). *TRADUCCIÓN AUTOMÁTICA NEURONAL Y TRADUCCIÓN AUTOMÁTICA ESTADÍSTICA: PERCEPCIÓN Y PRODUCTIVIDAD*. TRABAJO INÉDITO DE FIN DE MÁSTER. UNIVERSITAT AUTONOMA DE BARCELONA.

MARTÍN-MOR, A., Y HUERTA, RP (2017). "MTRADUMÁTICA Y LA FORMACIÓN DE TRADUCTORES EN TRADUCCIÓN AUTOMÁTICA ESTADÍSTICA". *REVISTA TRADUMÁTICA: TECNOLOGÍAS DE LA TRADUCCIÓN*, (15), PP. 97-115.

MARTÍNEZ, J. (2003). "EL CORPUS PARALELO: HERRAMIENTA PARA EL ESTUDIO DE TEXTOS PROCEDENTES DEL INGLÉS MODERNO TEMPRANO Y SUS TRADUCCIONES AL ESPAÑOL". *INTERLINGÜÍSTICA,* (14), PP. 719-728.

MOLINA, A. (2019). *CREACIÓN DE UN MOTOR DE TRADUCCIÓN AUTOMÁTICA ESTADÍSTICO (EN> SE) PARA TEXTOS DEL ÁMBITO FARMACÉUTICO. COMPARACIÓN CON OTROS MOTORES DE TRADUCCIÓN AUTOMÁTICA NEURONAL EXISTENTES*. TRABAJO INÉDITO DE FIN DE MÁSTER. UNIVERSITAT AUTONOMA DE BARCELONA.

PAPINENI, K., ROUKOS, S., WARD, T., Y ZHU, W. J. (2002). "BLEU: A METHOD FOR AUTOMATIC EVALUATION OF MACHINE TRANSLATION". *PROCEEDINGS OF THE 40TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, PP. 311-318.

PYM, A., Y TORRES-SIMÓN, E. (2021). *EFECTOS DE LA AUTOMATIZACIÓN EN LAS COMPETENCIAS BÁSICAS DEL TRADUCTOR: LA TRADUCCIÓN AUTOMÁTICA NEURONAL*. *OCUPACIONES Y LENGUAJE. INDICADORES Y ANÁLISIS DE COMPETENCIAS LINGÜÍSTICAS EN EL ÁMBITO LABORAL*, PP. 479-509. [DISPONIBLE EN: HTTPS://WWW.RESEARCHGATE.NET/PROFILE/ANTHONY-PYM-3/PUBLICATION/349255682_EFECTOS_DE_LA_AUTOMATIZACION_EN_LAS_COMPETENCIAS_BASICAS_DEL_TRADUCTOR_LA_TRADUCCION_AUTOMATICA_NEURONAL/LINKS/60B09291299BF13438F00420/EFECTOS-DE-LA-AUTOMATIZACION-EN-LAS-COMPETENCIAS-BASICAS-DEL-TRADUCTOR-LA-TRADUCCION-AUTOMATICA-NEURONAL.PDF]

POMAR, R. (2006). "EL ESPAÑOL EN BÉLGICA". *ENCICLOPEDIA DEL ESPAÑOL EN EL MUNDO: ANUARIO DEL INSTITUTO CERVANTES 2006-2007*, PP. 240-243. CÍRCULO DE LECTORES.

REI, R., STEWART, C., FARINHA, A. C., Y LAVIE, A. (2020). "COMET: A NEURAL FRAMEWORK FOR MT EVALUATION". ARXIV PREPRINT ARXIV:2009.09025.

SÁNCHEZ-MARTÍNEZ, F., SÁNCHEZ-CARTAGENA, V. M., PÉREZ-ORTIZ, J. A., FORCADA, M. L., ESPLA-GOMIS, M., SECKER, A., ... Y WALL, J. (2020). "AN ENGLISH-SWAHILI PARALLEL CORPUS AND ITS USE FOR NEURAL MACHINE TRANSLATION IN THE NEWS DOMAIN". *PROCEEDINGS OF THE 22ND ANNUAL CONFERENCE OF THE EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION,* PP. 299-308.

SCHWENK, H., CHAUDHARY, V., SUN, S., GONG, H., Y GUZMÁN, F. (2019). WIKIMATRIX: MINING 135M PARALLEL SENTENCES IN 1620 LANGUAGE PAIRS FROM WIKIPEDIA. ARXIV PREPRINT ARXIV:1907.05791.

TERTOOLEN, R. (2010). *EL ALCANCE DE LA TRADUCCIÓN AUTOMÁTICA, UN ESTUDIO DE LA TRADUCCIÓN AUTOMÁTICA DEL PAR DE LENGUAS ESPAÑOL NEERLANDÉS*. TRABAJO INÉDITO FIN DE MÁSTER. UNIVERSITEIT UTRECHT: UTRECH, PAÍSES BAJOS.

TIEDEMANN, J. (2012) "PARALLEL DATA, TOOLS AND INTERFACES IN OPUS". *PROCEEDINGS OF THE EIGHTH INTER-NATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION*(LREC-2012). PARIS: ELRA, 2214-8. [DISPONIBLE EN: WWW.LREC-CONF.ORG/PROCEEDINGS/LREC2012/PDF/463_PAPER.PDF].

TIEDEMANN, J. (2011) *BITEXT ALIGNMENT*. TORONTO: MORGAN & CLAYPOOL.

VALEMBOIS, V. (2007). "EL NEERLANDÉS, LA LENGUA DE MÁS DE VEINTE MILLONES DE EUROPEOS". *LETRAS*, (41), 171-205.

VANDEN BULCKE, P., Y DE GROOTE, C. (2016) JURIGENT, UN BANCO DE DATOS JURÍDICO NEERLANDÉS/ESPAÑOL DIFERENTE.

VERMEIREN, H. (2016). "INTERPRETAR DEL NEERLANDÉS (L1) AL ESPAÑOL (L4): PROPUESTAS PEDAGÓGICAS". *CLINA*, 2(2), 91-114.