Revista de Humanidades Digitales

Estudio del efecto de imprimación de la traducción automática sobre un corpus de textos del español institucional

Analysis of the priming effect of machine translation on Spanish texts in institutional contexts

Celia RICO PÉREZ
celrico@ucm.es
Universidad Complutense de Madrid
https://orcid.org/0000-0002-5056-8513

RESUMEN

Este artículo presenta un análisis del efecto de imprimación de la traducción automática en los textos institucionales de la Unión Europea traducidos al español. Se abordan dos preguntas clave: a) ¿es posible identificar alguna variación lingüística en los textos traducidos automáticamente coincidiendo temporalmente con los diferentes desarrollos de la tecnología de traducción automática?; b) si existen variaciones ¿hasta qué punto pueden deberse al efecto de imprimación de la traducción automática? Se trata de un estudio cuantitativo sobre cuatro aspectos: la diversidad léxica, la densidad léxica, el índice de la longitud del corpus (*lenght ratio*) y los patrones léxicos. Los resultados muestran ciertos indicios de imprimación de la traducción automática, aunque, como se indica en la conclusión, los datos no son concluyentes. Sería necesario complementarlos con un análisis cualitativo que examine casos individuales en contexto y que explore las variaciones lingüísticas que no se reflejan en los datos cuantitativos.

PALABRAS CLAVE

Imprimación de la traducción automática, diversidad y densidad léxica, corpus lenght ratio, corpus UCM-EUROPA.

ABSTRACT

This article presents an analysis of the priming effect of machine translation on institutional texts from the European Union translated into Spanish. Two key questions are addressed: a) Is it possible to identify any linguistic variation in machine-translated texts that temporally coincides with the different developments in machine translation technology?; b) If variations exist, to what extent can they be attributed to the priming effect of machine translation? This article presents a quantitative study focusing on four aspects: lexical diversity, lexical density, corpus length ratio, and lexical patterns. The results show some evidence of a machine translation priming effect. However, as indicated in the conclusion, the results are not conclusive. It would be necessary to complement them by exploring the corpus from a qualitative perspective, in order to examine individual cases in context, and analyze linguistic variations not reflected in the quantitative data.

KEYWORDS

Machine Translation Priming, Lexical Diversity and Density, Corpus Lenght Ratio, corpus UCM-EUROPA.

RHD 10 (2025)

ISSN
2531-1786

Revista de Humanidades Digitales http://revistas.uned.es/index.php/RHD/index
Recibido 20/07/2024 – Aceptado 30/01/2025

1. INTRODUCCIÓN

Con la llegada de la traducción automática neuronal y, más recientemente, con los sistemas de traducción basados en la arquitectura de transformadores (Vaswani et al., 2017), parece que ha aumentado la preocupación por saber hasta qué punto una lengua podría verse afectada como consecuencia, precisamente, del uso de estos sistemas de traducción artificial. Este efecto, denominado imprimación (*priming* en inglés) ha sido estudiado a lo largo de la última década en los trabajos de Lapshinova-Koltunski (2013 y 2015), Čulo y Nitzke (2016), Daems et al. (2017), Toral (2019) y Vanmassenhove et al. (2021), entre otros. Originalmente, el efecto de imprimación que reciben los textos traducidos se refiere al efecto psicológico por el cual la persona que traduce, al activar sus dos lenguas de trabajo simultáneamente, se ve influida por ambas a la hora de explorar y seleccionar los elementos potenciales del texto de llegada (Bangalore et al., 2016, p. 216), con la consecuencia de que algunos rasgos lingüísticos del texto original se *imprimen* en la traducción. Las características de la traducción humana se han investigado ampliamente tanto en los estudios de traducción como en la lingüística de corpus. En este campo, uno de los primeros estudios lo aporta Baker (1993) quien se refiere a las características universales de la traducción como ciertos rasgos lingüísticos que normalmente ocurren en los textos traducidos. Estos rasgos se concretan en un marcado aumento en el nivel de explicitación, desambiguación y simplificación, así como en la preferencia por estructuras gramaticales de la lengua de llegada. Con el avance de esta área de estudio, se han incorporado nuevos rasgos que caracterizan la imprimación de la traducción humana, tales como la influencia que ejerce el texto original (Teich, 2003), la hipótesis del ítem único (Tirkkonen-Condit, 2004) y la restricción del universal lingüístico (Kruger y van Rooy, 2016; Kajzer-Wietrzny y Ivaska, 2020).

A pesar de que la traducción automática se ha convertido en una parte integral de la traducción, el estudio de las características lingüísticas de la imprimación de esta tecnología es aún un área incipiente. Sin embargo, parece lógico pensar que cuando la traducción humana se apoya en herramientas tecnológicas estas puedan también contribuir al efecto de imprimación (Lapshinova-Koltunski, 2013 y 2015). En el caso de la traducción automática, la imprimación tendría su origen no en el efecto psicológico que experimenta la persona que traduce sino en el sesgo que aportan el algoritmo y los datos con los que se entrenan los propios sistemas de traducción.

Uno de los importantes usuarios de la traducción automática en Europa es la Comisión Europea, una institución que, por su propio mandato sobre el multilingüismo[1], se ve obligada a traducir grandes volúmenes de texto anualmente en las diferentes combinaciones lingüísticas de los países que componen la Unión Europea. En este proceso, la Dirección General de Traducción (DGT) de la Unión Europea suele recurrir a la traducción automática como herramienta de apoyo y ha sabido

[1] El multilingüismo está consagrado en la Carta de los Derechos Fundamentales de la Unión Europea (EUR-LEX 2012).

adaptarse a los diferentes desarrollos de esta tecnología a lo largo del tiempo, desde la primera implementación de un sistema de traducción basado en reglas en el año 1976 hasta la reciente incorporación de los modelos basados en inteligencia artificial. En este contexto, cabe preguntarse acerca del posible efecto de imprimación que las sucesivas incorporaciones de la traducción automática a lo largo del tiempo pueden haber tenido en las traducciones que se generan en el seno de la Comisión Europea. Así, este estudio se plantea con el objetivo de analizar hasta qué punto la integración de la traducción automática ha causado algún efecto de imprimación en los textos institucionales de la Unión Europea traducidos al español y descubrir, si este es el caso, qué rasgos lingüísticos caracterizan el citado efecto. A este respecto, conviene matizar que la investigación que aquí se presenta adopta un punto de vista descriptivo acerca "del lenguaje traducido automáticamente", en palabras de Jiménez Crespo (2023), con el ánimo de comprender los fenómenos de la lengua traducida artificialmente desde un marco de trabajo que deja atrás posturas generalizadoras y universalistas. Desde este punto de vista, nos alejamos de una conceptualización en la que el texto traducido supone una *desviación* de los rasgos lingüísticos que caracterizan el texto original, una postura que, de algún modo, afecta al estatus social de la persona que traduce y a la traducción como objeto cultural, lo cual se añade a la supuesta tendencia a la invisibilidad de la persona traductora y su falta de agencia en el proceso. En el caso de la traducción automática, cuando se analizan los textos traducidos por la máquina, podemos pensar que se contribuye a perpetuar el discurso sobre la falta de naturalidad y fluidez que se suele asociar a esta tecnología (Freitag et al., 2022) y que, de alguna manera, acaba afectando a las supuestas ineficiencias del sector profesional y la percepción un tanto negativa que de ello se desprende (Jiménez Crespo, 2023, p. 4).

Este artículo se estructura del siguiente modo. En primer lugar, se presentan los trabajos antecedentes sobre el efecto de la imprimación en traducción automática, seguidos de una revisión detallada de las diferentes implementaciones de esta tecnología a lo largo del tiempo en el seno de la Comisión Europea. Esta información permite establecer los hitos que marcan el periodo temporal al que se refiere el presente estudio y que se concreta, como veremos más adelante, entre los años 2006 y 2020. El artículo continúa con el apartado dedicado al diseño de la investigación y la metodología empleada, en el que se formulan los objetivos y preguntas de investigación, se describe el proceso de síntesis del corpus y se exponen los cuatro criterios de análisis empleados en el estudio, a saber: la diversidad léxica, la densidad léxica, el índice de la longitud del corpus (*lenght ratio*) y los patrones léxicos. En el siguiente apartado, resultados y discusión, se exponen los datos obtenidos tras la aplicación de cada criterio de análisis y se contrastan con otros estudios similares. El artículo se cierra con la conclusión de que es posible identificar ciertos indicios de variación lingüística en los textos traducidos automáticamente, aunque para afirmar que los datos sean concluyentes sería necesario abordar el estudio del corpus, de manera complementaria, desde una perspectiva cualitativa.

2. ANTECEDENTES

2.1. La traducción automática y el efecto de imprimación en los textos traducidos

Uno de los primeros estudios en el campo de estudio sobre la imprimación de la traducción automática lo encontramos en Daems et al. (2017), quienes evalúan las posibles interferencias de la traducción automática en aspectos léxicos, sintácticos y semánticos. Su estudio recoge dos tipos de datos. Por una parte, datos sobre la capacidad de un grupo de estudiantes para diferenciar entre los textos que se han traducido automáticamente y los que no y, por otra parte, datos estadísticos que permitan identificar los rasgos distintivos de esos mismos textos. En este último caso, se lleva a cabo un análisis exhaustivo según los siguientes índices: a) *type-token ratio* (TTR) como medida de la complejidad léxica del texto; b) el índice de perplejidad (*perplexity ratio*) del texto en comparación con un corpus de referencia y su normalización sobre la longitud del texto (*normalized perplexity*); c) el índice *term frequency-inverse document frequency* o tf-idf (Salton, 1989); d) el *Log Likelihood ratio* de todos los términos; e) el estudio de los rasgos sintácticos a partir del etiquetado de los textos y el cálculo de la frecuencia media de aparición de cada categoría de palabras (sustantivos, adjetivos, verbos, adverbios y preposiciones); y, por último, e) un análisis de los rasgos semánticos a partir de la identificación de listas de conectores como indicadores de la cohesión textual, calculando la media de frecuencia de aparición en el texto (Daems et al., 2017, p. 97). El resultado de este completo estudio es que no existen datos concluyentes que permitan determinar cuándo un texto se ha traducido automáticamente. Desde el punto de vista de la percepción del grupo de participantes en el estudio no fue posible distinguir entre una traducción humana y una automática y, de igual modo, los resultados del análisis computacional tampoco arrojaron datos concluyentes. En todo caso, los propios autores recomiendan cautela al interpretar sus resultados puesto que pueden estar condicionados por lo limitado del conjunto de datos que se emplearon, la combinación de idiomas y el género textual. De hecho, otros experimentos aportan datos que corroboran, precisamente, la idea contraria.

Este es el caso del trabajo de Čulo y Nitzke (2016), que analiza los patrones de variación terminológica en un conjunto de textos técnicos poseditados[2] del inglés al alemán en comparación con los traducidos de manera manual. Para realizar este estudio se emplea el coeficiente de perplejidad como medida de la variación terminológica que evidencia la imprimación de la traducción automática. Los resultados apuntan a que la traducción automática sí puede tener un efecto de imprimación en los textos. Es cierto que el corpus de estudio es muy reducido ya que se limita a tres textos técnicos y tres médicos con una longitud de 150 palabras cada uno, pero los propios autores, al reconocer esta limitación, avanzan la hipótesis de cuál puede ser el resultado en corpus más

[2] La posedición es la corrección que realiza un traductor profesional de una traducción *en bruto* generada previamente por un programa de traducción automática (ISO, 2017).

grandes, con otras combinaciones lingüísticas y empleando diversos programas de traducción automática. En un estudio posterior, Toral (2019) recoge esta propuesta y plantea un análisis computacional más ambicioso. En su caso, se recoge un subconjunto de traducciones de diferentes corpus: Taraxü (Avramidis et al., 2014), IWSLT (Cettolo et al., 2015) y Microsoft Human Parity (Hassan et al., 2018). Estos conjuntos de datos contienen traducciones humanas y poseditadas, diferentes combinaciones lingüísticas entre inglés, alemán, francés y chino, así como varias áreas de especialidad. Además, al ser datos traducidos por diferentes programas de traducción automática a lo largo de una década permiten llevar a cabo un análisis longitudinal (Toral, 2019, p. 275). Sus conclusiones apuntan a que sí hay ciertos rasgos que indican la imprimación en los textos traducidos mediante traducción automática comparados con las características de los textos traducidos de manera manual. En concreto, los rasgos que se observan son los siguientes: variedad y densidad léxica menor, longitud similar a la del texto original y secuencias gramaticales también similares a las del texto original (Toral, 2019, p. 278). Más recientemente, Vanmassenhove et al. (2021) analizan los resultados de la traducción automática de un sistema neuronal en las combinaciones lingüísticas con inglés, francés y español. Los criterios de análisis corresponden a la riqueza léxica y la diversidad gramatical y, en ambos casos, se constata lo que los autores denominan un *lenguaje artificialmente empobrecido* (Vanmassenhove et al., 2021, p. 2203), con signos de interferencia del texto original e imprimación de la traducción automática. A estos datos, se une el estudio llevado a cabo por Niu y Jiang (2024) con un alcance mucho mayor que estudios previos puesto que se centran en un corpus comparable de un total de 593 578 palabras, distribuidas en tres géneros diferentes: literatura contemporánea, documentos gubernamentales y resúmenes académicos (Niu y Jiang, 2024, p. 4). Su análisis muestra los rasgos de imprimación en la simplificación de la traducción automática para la combinación lingüística chino-inglés, concretamente en los rasgos de variedad y densidad léxica.

A la vista de todos estos trabajos, parece que hay una serie de indicios que caracterizan la imprimación de la traducción automática y que se concretan en los siguientes: diversidad y densidad léxica menor, longitud similar a la del texto original y secuencias gramaticales también similares a las del texto original. A este respecto, el estudio que se presenta en este artículo parte de la idea de que la traducción automática aporta una imprimación al texto traducido. En este sentido, esta propuesta explora hasta qué punto esto puede ser cierto en los textos que se han traducido al español en la Comisión Europea a lo largo del tiempo y por efecto de los diferentes sistemas de traducción automática que se han ido implementando.

Como paso previo al análisis de este efecto de imprimación, en el siguiente apartado se presenta un recorrido temporal por las sucesivas implementaciones de la traducción automática en la Comisión Europea. Este recorrido servirá para establecer los hitos que componen el periodo al que se refiere este estudio.

2.2. La traducción automática en la Comisión Europea

La primera implementación de la traducción automática en el seno de la Comisión Europea fue el programa Systran, utilizado inicialmente en 1976 en el Servicio de Traducción (SdT) de la que entonces era la Comisión de las Comunidades Europeas. Este sistema estaba basado en reglas y había sido desarrollado por Peter Toma en la década de 1960 para las fuerzas aéreas del ejército de Estados Unidos (Toma, 1976). Coincidiendo en el tiempo con el uso de Systran, y tras constatar la Comisión la complejidad que suponía adaptar este sistema de traducción al conjunto de las lenguas europeas (Celex, 1994, pp. 32), se creó el programa de investigación Eurotra, que se extendió desde 1983 a 1990 (Maegaard y Perschke, 1991). Eurotra planteaba dos objetivos: por una parte, el desarrollo de un prototipo pre-industrial de un sistema de traducción automática para todas las lenguas de la Comunidad Europea (nueve, en ese momento) y, por otra, la creación de conocimientos científicos en este campo en toda Europa. El primer objetivo no llegó a cumplirse en toda su magnitud[3], aunque sí hubo múltiples desarrollos experimentales que se llevaron a cabo en los diferentes centros de investigación que participaban (Maegaard, 1995). En cuanto al segundo, sí puede afirmarse que se llevaron a cabo importantes investigaciones que contribuyeron a que Europa se convirtiera en una referencia en esta área de estudio con la creación de grupos de trabajo en lingüística computacional (Celex, 1994, pp. 3, 5)[4].

A pesar de que el programa Eurotra había quedado cerrado y de que Systran planteaba problemas para su adaptación, el interés por la traducción automática seguía vivo, sobre todo, por la necesidad de afrontar las tareas de traducción, que para la Comisión suponían un gasto económico importante (Celex, 1994, p. 117). Con todo, Systran continuó siendo el programa utilizado en el SdT de la Comisión a lo largo de las dos décadas siguientes, con adaptaciones y mejoras a cargo de la DGXIII (EUR-Lex, 1997). Con el desarrollo de los primeros programas para la gestión de memorias de traducción en la década de 1990 (Rothwell et al., 2023), el SdT inicia una nueva etapa en el uso de esta tecnología, incorporándola junto a la traducción automática. El hecho de que se empleasen ambas tecnologías es relevante porque, a partir de ese momento, la Comisión Europea crea su propio ecosistema de traducción compuesto por diferentes herramientas. En el caso de la traducción automática, el sistema utilizado recibió el nombre de ECMT (*European Commission Machine Translation*), que es una evolución del programa Systran original, basado en reglas (European Commission, 2008).

Con la irrupción de la traducción automática de base estadística como consecuencia, principalmente, del desarrollo del motor de código abierto Moses (Koehn et al., 2007), la DGT se une también a estos avances y en 2010 los incorpora a su nuevo sistema de traducción MT@EC, que

[3] Es necesario tener en cuenta que este objetivo suponía la creación de un sistema de traducción automática para las nueve lenguas oficiales de la Comunidad Europea (español, danés, alemán, griego, inglés, francés, holandés y portugués) en todas sus combinaciones, es decir, un total de 72 pares de idiomas (Celex, 1994, p. 33).

[4] En el repositorio digital de la *Association for Computer Machinery* (ACL) se pueden consultar todas las publicaciones de la época referidas al proyecto con la palabra clave "Eurotra": https://aclanthology.org/.

combina la tecnología estadística con la base de datos de documentos traducidos de la DGT (Euramis) en las 23 lenguas oficiales de la UE (Foti, 2012). Este volumen de datos supuso que MT@EC se entrenase con más de 1.000 millones de frases en las 24 lenguas oficiales de la UE, producidas por los traductores de las instituciones de la UE durante las décadas previas (Steinberger et al., 2014). A partir de este momento, el sistema MT@EC se convierte en una parte clave del proceso de traducción en la DGT junto a otras herramientas de apoyo a la traducción como son la base de datos terminológica IATE (*InterActive Terminology for Europe*), un sistema de reconocimiento de voz y el sistema de gestión de memorias de traducción Euramis. En el año 2016, el sistema MT@EC ofrecía la traducción de un total de 552 pares de lenguas, de los cuales 76 eran combinaciones directas y el resto se traducía con el inglés como lengua intermedia. El volumen estimado de traducción es de 100.000 páginas al día, con picos de hasta 400.000 páginas en determinados momentos (European Commission, 2016, p. 12). Al año siguiente, en 2017, cuando irrumpe en el mercado Google con los modelos de traducción neuronales (Sánchez Ramos y Rico Pérez, 2020, p. 6), el compromiso de la Comisión Europea con la traducción le lleva a adoptar también estos modelos y a desarrollar su propio sistema de traducción automática neuronal e*Translation*, que queda incorporado en el flujo de trabajo de la DGT junto al resto de herramientas de traducción, en sustitución de MT@EC (Translation Centre, 2019, p. 3). En su configuración actual, es de suponer que e*Translation* incorpore los avances que se han producido en inteligencia artificial con los grandes modelos de lenguaje (Brown et al., 2020), aunque la documentación al respecto no parece estar disponible. Sí podemos encontrar información acerca de diferentes acciones que ha llevado a cabo el Centro de Traducción para ampliar la cobertura de dominio de los motores personalizados integrados neuronales de su sistema de traducción automática, concretamente en dos nuevos ámbitos: finanzas y asuntos sociales (Translation Centre, 2023, p. 18). En este sentido, se han construido dos sistemas piloto por cada ámbito para las combinaciones (inglés-francés e inglés-sueco), que aprovechan, igualmente, el ingente volumen de datos traducidos de la base de datos Euramis. A lo largo del año 2023, el Centro de Traducción ha puesto en marcha una estrategia multimotor para la traducción automática. Esto significa que "utilizará varios sistemas de traducción automática al mismo tiempo y proporcionará a los traductores el mejor resultado gracias a un sistema automático de puntuación de la calidad que determinará la traducción automática más adecuada para cada frase" (Translation Centre, 2023, p. 41).

A modo de resumen, se presentan en la Figura 1 los diferentes hitos de la traducción automática en la Comisión Europea a lo largo del tiempo:

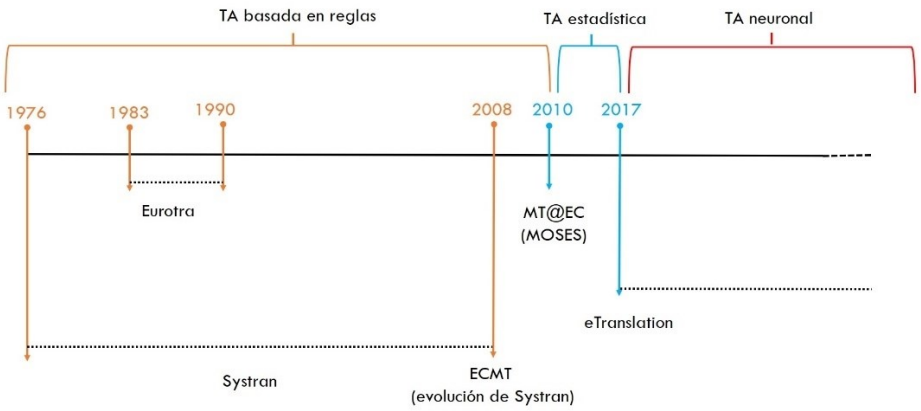


Figura 1. Evolución de los sistemas de traducción automática en la Comisión Europea (1976-actualidad): de los enfoques basados en reglas a la traducción neuronal. Fuente: elaboración propia.

3. DISEÑO DE LA INVESTIGACIÓN Y METODOLOGÍA DE ANÁLISIS

3.1. Objetivos

El trabajo que se presenta en este artículo tiene como objetivo analizar el posible efecto de imprimación que ha generado el uso de la traducción automática a lo largo del tiempo en los textos institucionales de la Unión Europea traducidos al español y coincidiendo con los hitos en la implementación de la traducción automática. Con el fin de alcanzar el objetivo, se plantean dos preguntas de investigación:

a) ¿Es posible identificar algún tipo de variación lingüística en los textos traducidos automáticamente en el periodo 2006-2020, coincidiendo con las diferentes implementaciones de la tecnología?
b) Si existen estas variaciones ¿hasta qué punto estas pueden tener su origen en el efecto de imprimación de la traducción automática?

Para dar respuesta a estas preguntas, se plantea un análisis diacrónico de los textos traducidos al español en el seno de la Dirección General de Traducción (DGT) de la Comisión Europea, con el fin de evaluar el posible efecto que haya podido tener el uso de la traducción automática.

3.2. Corpus de análisis: el corpus UCM-EUROPA

El estudio que aquí se presenta abarca el periodo 2006-2020, que corresponde al periodo en el que existen datos públicos disponibles de las traducciones realizadas en la Comisión Europea. Precisamente es en 2006 cuando, por primera vez, la DGT de la Comisión Europea pone a disposición pública las memorias de traducción que contienen el conjunto de datos de la legislación

europea dentro del *Acquis Communautaire*[5] en los 24 idiomas de la Unión y que se han ido actualizando anualmente a partir de 2011 (Steinberger et al., 2012 y 2014). La relevancia de estas memorias de traducción está en que permiten observar de primera mano las características de la lengua traducida puesto que, como hemos visto en el apartado 2.2., son el producto de la interacción de las diferentes herramientas de traducción, entre las que se encuentra la traducción automática. Las memorias de traducción de la DGT están disponibles en el portal de recursos de tecnología de la lengua (DGT-Translation Memory, 2024), almacenadas en formato .tmx.

Para crear el corpus de estudio, corpus UCM-EUROPA (Rico Pérez, 2024), se descargaron los datos de las traducciones al español, correspondientes a los diferentes hitos de la traducción automática según se describe en la Figura 1, comenzando por 2006 (el primer año en el que están disponibles las memorias de traducción) y hasta 2020 (el último año del que se disponen datos). Para cada año/hito se extrajo un subcorpus de alrededor de 500000 tokens por idioma, según puede verse en la Tabla 1.

año	idioma	unidades de traducción	types	tokens
2006	ES	30493	27387,00	478972,00
2006	EN		23746,00	428542,00
2008	ES	21575	25511,00	487282,00
2008	EN		19074,00	433634,00
2010	ES	21902	23791,00	480970,00
2010	EN		18651,00	426745,00
2017	ES	21479	19477,00	486512,00
2017	EN		16082,00	441864,00
2020	ES	23252	18028,00	490960,00
2020	EN		14725,00	417778,00

Tabla 1. Corpus de estudio. Fuente: elaboración propia.

Para extraer cada subcorpus, los archivos .tmx correspondientes a cada memoria se procesaron con el programa de gestión de memorias *Phrase*[6] en su versión de uso académico y se generaron los correspondientes archivos .txt para cada par de idiomas. Posteriormente, cada uno de los archivos se etiquetó con información gramatical con el programa TagAnt (Anthony, 2022 y 2023), que utiliza los modelos de lengua pre-instalados de SpaCy. El corpus UCM-EUROPA está disponible en Rico Pérez (2024).

[5] El *Acquis Communautaire* es el conjunto normativo vigente en la Unión Europea (UE). Comprende las normas originarias contenidas en los tratados fundacionales o en sus modificaciones, la legislación derivada dictada para el desarrollo de los tratados, las declaraciones y resoluciones dictadas por los organismos europeos y también los tratados internacionales suscritos por la UE. El *Acquis* puede consultarse en este enlace: https://eur-lex.europa.eu/ES/legal-content/glossary/acquis.html.

[6] El programa *Phrase* está disponible en este enlace: https://eu.phrase.com/idm-ui/signin.

3.3. Criterios de análisis

Para llevar a cabo el análisis de los datos se seleccionaron cuatro criterios que permiten explorar posibles indicios del efecto de imprimación de la traducción automática. Los criterios son los siguientes:

a) Diversidad léxica. Para estudiar este parámetro se emplean dos métricas: *type-token ratio* (TTR) y *measure of textual lexical diversity* (MTLD). En el primer caso, la medida TTR se define como la relación entre el número de palabras distintas (*type*) y el número total de palabras en el corpus (*token*). Esta relación nos permite hacer el recuento de las palabras distintas en el corpus, de manera que cuanto más alto es el índice TTR mayor es el grado de diversidad léxica (Stefanowitsch, 2020, pp. 315-316). La diversidad léxica se expresa mediante la siguiente fórmula:

$$TTR = \frac{número\ de\ types}{número\ de\ tokens}$$

La medida TTR ha recibido críticas dada su fuerte correlación negativa con la longitud del texto[7] por lo que estudiamos también la medida MTLD al ser menos sensible a la longitud del texto. La medida MTLD se evalúa de manera secuencial como la longitud media de cadenas secuenciales de palabras en un texto que mantiene un valor TTR determinado (McCarthy y Jarvis, 2010, p. 384).

b) Densidad léxica, definida como la relación entre las palabras con contenido semántico (sustantivos, adjetivos, verbos y adverbios) y el número total de tokens. Esta relación permite conocer cuánta información contiene un texto, de manera que un texto con mayor cantidad de palabras con contenido semántico es un indicio de que este contiene más información que un texto con mayor proporción de palabras funcionales (preposiciones, interjecciones, pronombres, conjunciones) (Johanson, 2008, pp. 61-65). La densidad léxica se expresa con la siguiente fórmula:

$$densidad\ léxica = \frac{palabras\ contenido}{tokens}$$

[7] El TTR desciende cuando aumenta la longitud del texto, como documentan Richards (2006) o McCarthy y Jarvis (2010).

c) Índice de la longitud del corpus (*length ratio*). En este caso, se trata de calcular la diferencia absoluta entre la longitud del corpus original en inglés (medida en número total de tokens) y la longitud del corpus traducido al español (medida en número total de tokens), normalizado a la longitud del corpus original en inglés.

$$lenght\ ratio = \frac{tokens\ corpus\ original - tokens\ corpus\ traducido}{tokens\ corpus\ original}$$

Con este dato podremos analizar qué relación guarda la longitud de texto traducido automáticamente con el texto original. Por lo general, los sistemas de traducción automática tienden a producir oraciones de longitud similar al texto original al no distanciarse de los patrones lingüísticos originales. Cuando esto ocurre, el *length ratio* tiende a ser bajo (Aranberri, 2020).

d) Patrones léxicos. En este caso se han estudiado los patrones léxicos que generalmente presentan los términos complejos (Sánchez-Saus, 2022, p. 81):

- sustantivo + adjetivo
- sustantivo + de + sustantivo
- adjetivo + sustantivo + de + sustantivo

El estudio de estos patrones nos permitirá conocer la variación léxica de las soluciones aportadas por la traducción automática (Čulo y Nitzke, 2016). Los patrones se han identificado mediante la opción *n-gram* de AntConc y el uso de expresiones regulares, aplicando posteriormente el índice TTR a la relación *n-gram types/n-gram tokens,* de manera que podamos conocer la variación léxica de los patrones terminológicos, esto es, cuántos de ellos son únicos (*types*) con respecto al total (*tokens*).

El análisis de estos cuatro parámetros nos permitirá identificar posibles indicios de imprimación de la traducción automática y abordar las preguntas de investigación según se ha presentado en el apartado 3.1.

4. RESULTADOS Y DISCUSIÓN

Se detallan a continuación los resultados del análisis, clasificados según los cuatro criterios establecidos.

4.1. Diversidad léxica

Como se indica más arriba, la diversidad léxica nos permite conocer cómo evoluciona el uso del léxico a lo largo del periodo 2006-2020. En la Tabla 2 se muestran los datos obtenidos al analizar las medidas TTR y MTLD[8].

Año	Idioma	Types	Tokens	TTR	MTLD
2006	Español	27387,00	478972,00	0,057	645,29
2008	Español	25511,00	487282,00	0,052	608,50
2010	Español	23791,00	480970,00	0,049	687,65
2017	Español	19477,00	486512,00	0,040	472,01
2020	Español	18028,00	490960,00	0,037	394,06

Tabla 2. Diversidad léxica en el corpus de traducciones.
Fuente: elaboración propia.

Como puede observarse, los valores obtenidos tanto en el índice TTR como en MTLD descienden a lo largo de los años y pasan, en el caso del TTR de un valor de 0,057 en el año 2006 al valor de 0,037 en el año 2020. En el caso de la medida MTLD, el valor inicial en 2006 es de 645,29 y desciende a 394,06 en 2020. En los gráficos 1 y 2 se muestran estas tendencias.

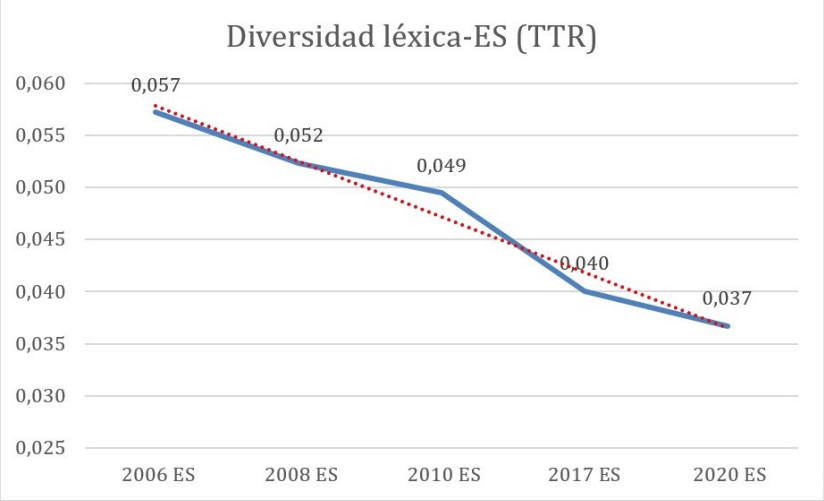


Gráfico 1. Tendencia de la diversidad léxica (TTR).
Fuente: elaboración propia.

[8] Para el cálculo de la medida MTLD se ha empleado la librería de Kyle (2020).

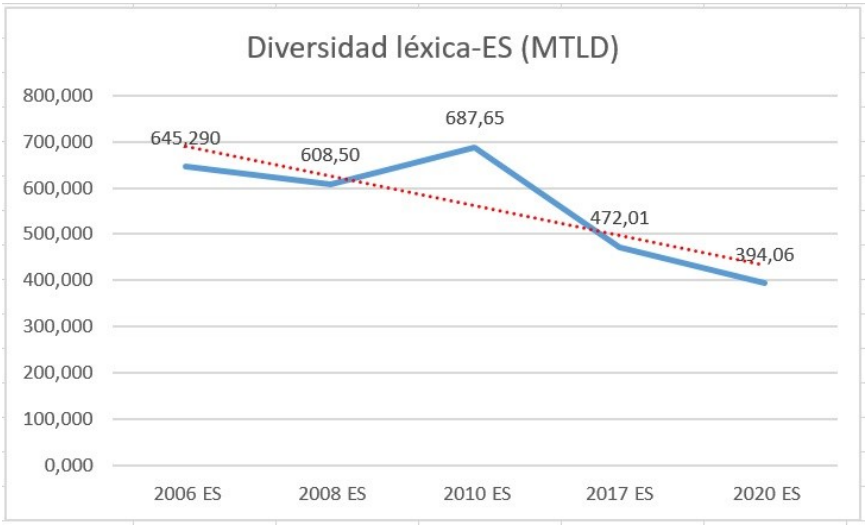


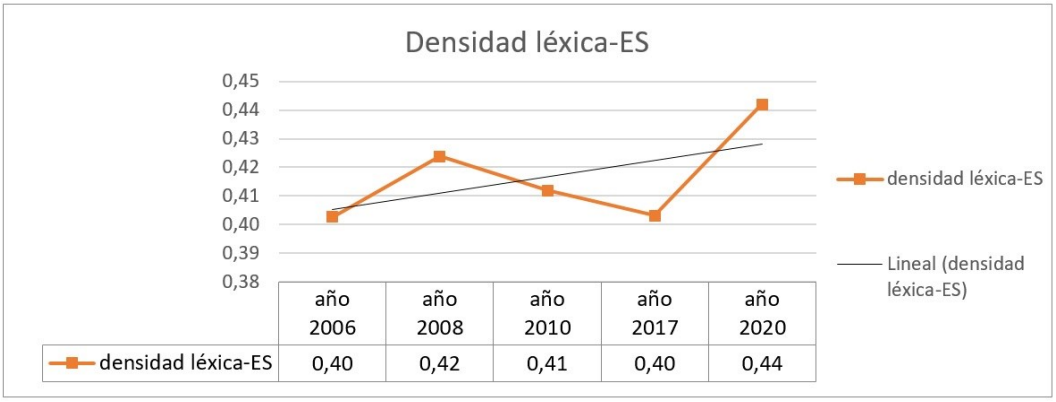
Gráfico 2. Tendencia de la diversidad léxica (MTLD).
Fuente: elaboración propia

Es importante puntualizar que, puesto que el corpus de 2020 tiene 11988 tokens más que el de 2006, cabría pensar que la diferencia de TTR puede deberse en parte a la diferencia de longitud. El TTR baja de 0,057 a 0,052 entre 2006 y 2008, y aunque la longitud del corpus desciende en 8310 tokens, la relación entre las dos magnitudes no es linear (Richards, 2006). Si examinamos la correlación entre las dos series de la Tabla 2 (tokens y TTR) con las funciones *spearmanr* y *pearsonr* del paquete scipy. stats de Python, vemos que la correlación (de -0,7 y -0,78, respectivamente) no es estadísticamente significativa ($p > 0,05$) lo que podría apuntar a que las limitaciones de TTR no afectan negativamente a este estudio.

Recordemos que en el año 2006 se estaba utilizando en la DGT el programa Systran, un sistema de traducción basado en reglas que, combinado con las memorias de traducción, se había usado desde 1976 con sucesivas actualizaciones. En 2008 encontramos el siguiente hito en la evolución de la traducción automática en la DGT con la implementación del programa ECMT y, en este momento, los datos indican un ligero descenso en la diversidad léxica del corpus estudiado. Este descenso continúa con la incorporación, en 2010, de Moses (basado en tecnología estadística), hasta llegar al año 2020, cuando se implanta e*Translation* (tecnología de redes neuronales). Los datos recogidos parecen apuntar a que, con el transcurso del tiempo, se ha producido un efecto de simplificación léxica puesto que el vocabulario que se obtiene en la traducción es menos diverso. Para Vanmassenhove et al. (2021, p.2203) esta tendencia descendente indica un lenguaje artificialmente empobrecido que, en el caso de la traducción automática neuronal, puede tener su origen en el conjunto de datos de entrenamiento, de manera que cuando el sistema de traducción encuentra un texto que no está reflejado en estos datos, se obtiene una traducción lineal del texto original (DataLit^{MT}, 2023, p. 6). Esta simplificación léxica por efecto de la traducción automática parece un fenómeno común que se da también en los estudios realizados por Lapshinova-Koltunski (2015), Čulo y Nitzke (2016), Toral (2019) y Aranberri (2020).

4.2. Densidad léxica

Veamos, a continuación, los datos que se desprenden del estudio de la densidad léxica en el corpus de traducciones. Recordemos que, en este caso, se trata de medir la relación entre las palabras con contenido semántico (sustantivos, adjetivos, verbos y adverbios) y el número total de *tokens* del corpus. En el Gráfico 3 podemos ver cómo a lo largo del periodo de estudio se da una tendencia ascendente en la densidad léxica, es decir, que el número de palabras con contenido semántico en relación al total del corpus aumenta desde el primer año de estudio (2006) hasta el último que se toma en consideración (2020).



	año 2006	año 2008	año 2010	año 2017	año 2020
densidad léxica-ES	0,40	0,42	0,41	0,40	0,44

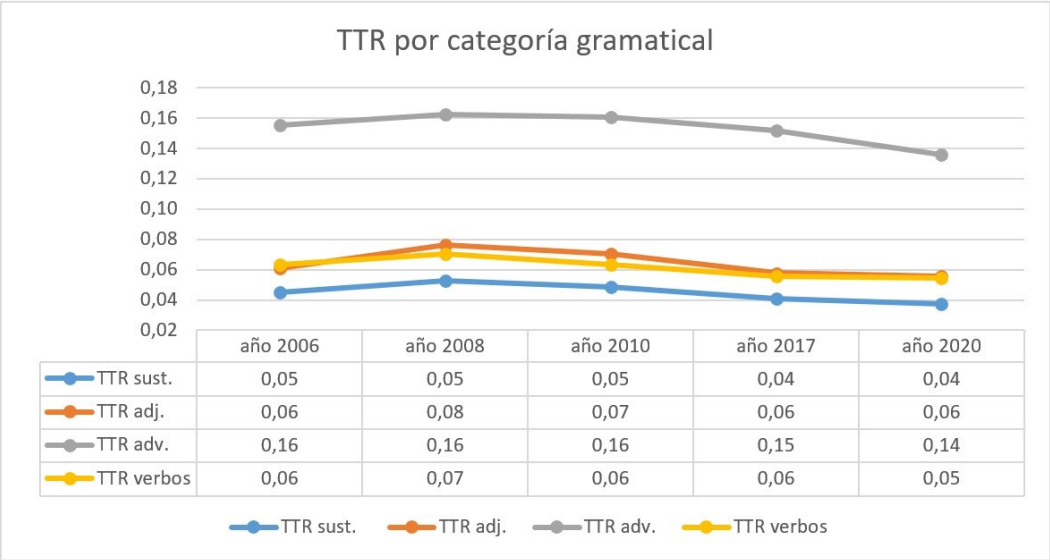
Gráfico 3. Tendencia de la densidad léxica agrupada. Fuente: elaboración propia

En este caso, podría afirmarse, *a priori*, que no existe una correlación entre la densidad léxica y los datos que hemos visto en el apartado anterior al analizar la diversidad léxica, que indican una tendencia descendente. Parece, entonces, que hay una contradicción en los resultados que se obtienen entre uno y otro índice. Sin embargo, tal como apunta Johansson (2008), esta situación es posible:

> It is, however, theoretically possible that a text has high lexical diversity (i.e. contains many different word types), but low lexical density (i.e. contains many pronouns and auxiliaries rather than nouns and lexical verbs), or, vice versa, that a text has low lexical diversity .e. the same words or phrases are repea ed over and over) but high lexical density (i.e. the words that are repeated are nouns, adjective or verbs) (Johanson, 2008, p. 61).

Según vemos, puede ocurrir entonces que en nuestro corpus de análisis se dé precisamente la circunstancia de que la *diversidad léxica* sea baja (es decir, que el corpus contenga un conjunto de palabras que se repiten una y otra vez) y que la *densidad léxica* sea alta (es decir, que las palabras que se repiten sean sustantivos, adjetivos o verbos). Con el fin de determinar si esto es así, estudiamos el índice TTR de las diferentes categorías gramaticales. Para ello se identifican en primer lugar los n-gramas de cada categoría con la opción n-gram de AntConc y el uso de expresiones regulares. Posteriormente se identifican el número de n-gram types y de n-gram tokens para cada categoría y se aplica el índice TTR a la relación *n-gram types/n-gram tokens*. De este modo,

podemos conocer la diversidad léxica de cada categoría gramatical y saber, para cada una de ellas, cuántos son únicos (*types*) con respecto al total (*tokens*). En el Gráfico 4 se muestran los resultados.



	año 2006	año 2008	año 2010	año 2017	año 2020
TTR sust.	0,05	0,05	0,05	0,04	0,04
TTR adj.	0,06	0,08	0,07	0,06	0,06
TTR adv.	0,16	0,16	0,16	0,15	0,14
TTR verbos	0,06	0,07	0,06	0,06	0,05

Gráfico 4. TTR por categoría gramatical. Fuente: elaboración propia

A partir de estos datos, se observa que la diversidad léxica en el uso de palabras con contenido semántico tiende a descender ligeramente a lo largo del periodo analizado y, aunque las diferencias son muy pequeñas (0,01 puntos) podríamos hablar, con cierta cautela, de una tímida tendencia a la simplificación.

4.3. Índice de la longitud del corpus (*length ratio*)

El análisis de la longitud del corpus es interesante porque nos permite ver hasta qué punto el uso de la traducción automática influye en que una traducción pueda tener (o no) una longitud similar al texto original. La hipótesis que se maneja, en este caso, es que una traducción humana tiende a ser más libre en cuanto a la longitud del texto en comparación con el original, mientras que una traducción automática muestra una longitud similar (con un índice menor) (Toral, 2019). En el Gráfico 5 se muestra el resultado de este análisis.

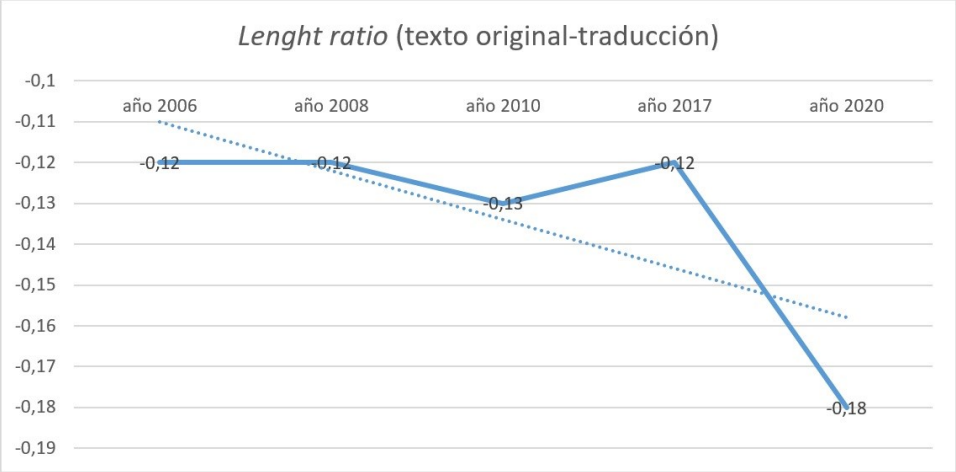


Gráfico 5. Análisis del *length ratio*. Fuente: elaboración propia.

Según vemos, en los dos primeros años del periodo objeto de estudio, la longitud se mantiene constante, pero en el año 2010, coincidiendo con la introducción de Moses como sistema de traducción automática, se da una caída en este índice. Este descenso puede explicarse, tal como apunta Lapshinova-Koltunski, E. (2015, p. 101), por las propias restricciones léxicas del sistema estadístico Moses, que, al depender de datos paralelos para su entrenamiento, deja sin traducir aquellas palabras para las que no encuentra una traducción.

En 2017, el índice asciende a 0,12 (es el momento en que se adopta el sistema e*Translation*), pero cae de manera importante en 2020 cuando se empieza a usar la traducción automática neuronal. En este caso, es relevante puntualizar que, tal como apunta Aranberri (2020, p.97), los sistemas de traducción automática tienden a producir traducciones con una longitud similar a la del texto original porque no tienen la capacidad para distanciarse de los patrones del original. Cuando esto ocurre, el *length ratio* tiende a ser bajo. Es necesario, además, tener en cuenta las diferencias de longitud entre la traducción y el texto original pueden deberse a una traducción incorrecta o incompleta.

4.4. Patrones léxicos

Con el análisis de los patrones léxicos estudiamos si se ha producido alguna variación en las soluciones aportadas por la traducción automática a lo largo del tiempo y para los hitos tecnológicos seleccionados. Como apuntamos anteriormente, las estructuras objeto de análisis son las siguientes:

- sustantivo + adjetivo (Gráfico 6).
- sustantivo + de + sustantivo (Gráfico 7)
- adjetivo + sustantivo + de + sustantivo (Gráfico 8)

En el caso del *patrón sustantivo + adjetivo*, el Gráfico 6 muestra una tendencia descendente que pasa de 0,124 en el año 2006 (traducción automática basada en reglas) a 0,110 en 2020 (traducción automática neuronal). Si bien es cierto que el índice asciende en el paso del año 2006 a 2008, eso no impide el descenso en el resto del periodo, tal como ya hemos apuntado. A pesar de que las diferencias no son realmente significativas (menos de 0.1 puntos en todos los casos), sí parece relevante la tendencia descendente a la que apuntan y que podría ser indicativa de una variación originada por el uso de la traducción automática, como discutiremos más adelante.

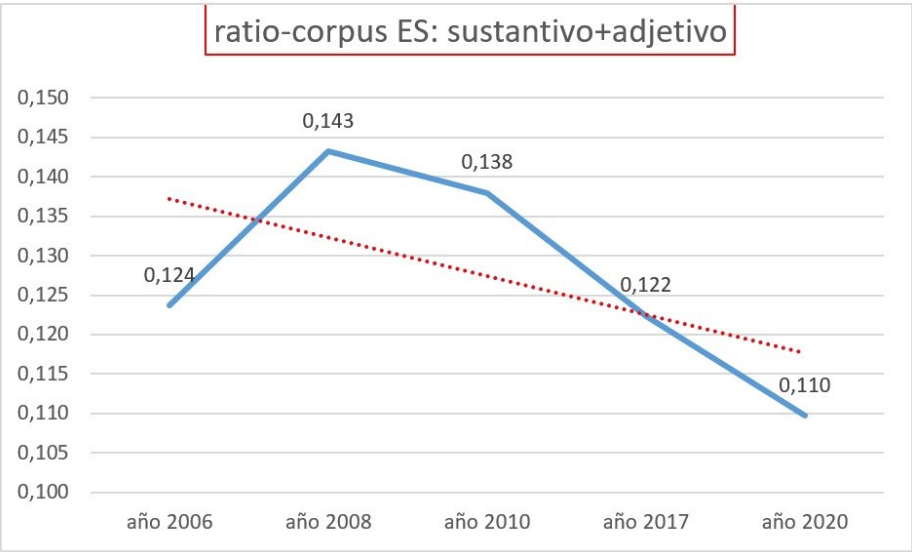


Gráfico 6. Patrón sustantivo + adjetivo. Fuente: elaboración propia.

En el patrón *sustantivo + de + sustantivo*, el Gráfico 7 muestra también una tendencia descendente, aunque, en este caso, se da un incremento en el índice entre los años 2006 y 2010, coincidiendo, de nuevo, en 2008 con la incorporación del sistema de traducción automática ECMT, cuando el índice se incrementa a 0,148 y en 2010 con el sistema Moses, con un índice de 0,154. En todo caso, la tendencia del periodo completo es descendente.

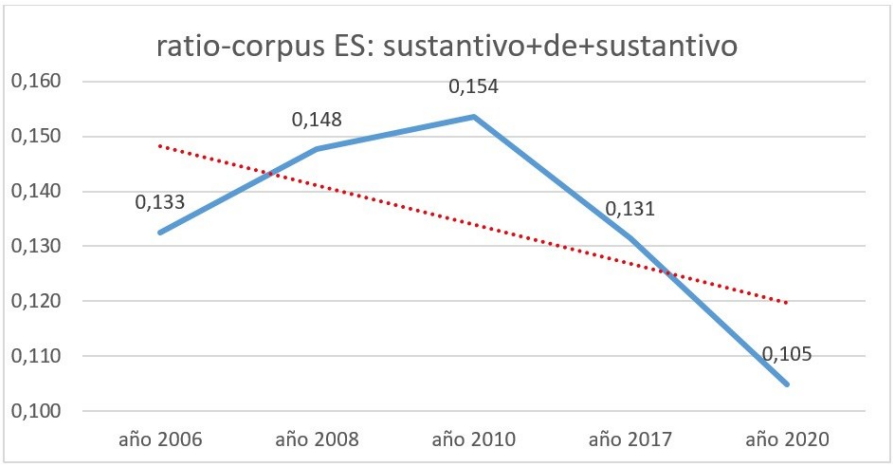


Gráfico 7. Patrón sustantivo + de + sustantivo. Fuente: elaboración propia.

Por último, el patrón *adjetivo* + *sustantivo* + *de* + *sustantivo*, que se muestra en el Gráfico 8, indica un ligero ascenso que pasa de 0,635 en el año 2006 (con Systran, traducción automática basada en reglas) a 0,702 en el año 2020. Es interesante mencionar el descenso que se da en el año 2017, con la introducción de *eTranslation*, basado en tecnología neuronal. El descenso en el patrón podría deberse al efecto de simplificación observado en estudios como el que plantean, por ejemplo, Niu y Jian (2024: p. 1), con resultados que muestran que la traducción automática aporta soluciones más simplificadas que la traducción humana. De manera complementaria, otra posible causa podría estar en el reemplazado de los citados patrones por expresiones más variadas, más idiomáticas, que hacen menos necesaria la repetición del patrón *adjetivo* + *sustantivo* + *de* + *sustantivo*. De hecho, en las primeras etapas de la traducción automática neuronal industrial (en 2016), se llevaron a cabo evaluaciones que mostraban que la traducción neuronal producía resultados más fluidos que los estadísticos (Toral y Sánchez-Cartagena, 2017; o Bojar et al., 2016). En todo caso, sería necesario disponer de un análisis cualitativo de ejemplos para corroborar esta observación[9].

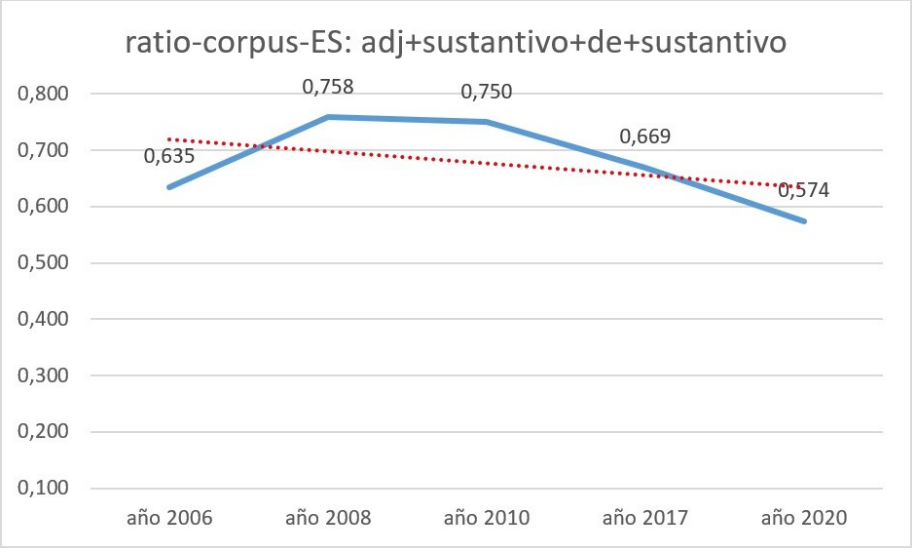


Gráfico 8. Patrón adjetivo + sustantivo + de + sustantivo.
Fuente: elaboración propia

A la vista de los datos obtenidos, parece que hay un efecto de imprimación en los textos traducidos automáticamente, que tienden a disminuir la diversidad de los recursos lingüísticos adoptados y apuntan a una tendencia en la variación de los patrones léxicos. Según destacan va-

[9] A este respecto es interesante el análisis que lleva a cabo Krüger (2020: 213) en el que aporta datos cualitativos sobre el fenómeno de explicitación de la traducción automática en un sistema de traducción neuronal. Por una parte, muestra que el sistema parece haber aprendido ciertos giros de explicitación a partir de los datos de entrenamiento, que le permiten resolver ciertas ambigüedades en el plano sintáctico del texto original. Por otra parte, el sistema llega a insertar unidades léxicas autónomas en el texto traducido cuando no existe un desencadenante lingüístico explícito en el texto original que provoque tal cambio. A pesar de que los sistemas de traducción automática neuronal pueden aprovechar información contextual más rica que las arquitecturas anteriores, este "aprovechamiento" sigue siendo fragmentario. En el caso que nos ocupa, con respecto a la simplificación, sería necesario abordar el fenómeno de manera cualitativa de un modo similar al plateado por Krüger (2020).

rios estudios (Martikainen y Kubler, 2016; Čulo y Nitzke, 2016; Farrel, 2018), el empleo de la traducción automática puede llevar, incluso, a una sobrerrepresentación de ciertas soluciones de traducción que son más frecuentes en comparación con la traducción humana. Se trataría, por tanto, de un fenómeno de homogenización como resultado del efecto de imprimación de la traducción automática, que tiende a favorecer las traducciones que se dan con mayor frecuencia en los datos de entrenamiento (Volkart y Bouillon, 2022, p. 76).

5. CONCLUSIONES

El estudio presentado en este artículo tiene como objetivo analizar el posible efecto de imprimación de la traducción automática a lo largo del tiempo en los textos institucionales de la Unión Europea traducidos al español y coincidiendo con los diferentes hitos tecnológicos en el uso de la traducción automática en la DGT: Systran y ECMT (sistemas basados en reglas), Moses (sistema estadístico) y *eTranslation* (sistema basado en redes neuronales). Para abordar este objetivo, se han planteado dos preguntas de investigación. La primera se centra en descubrir hasta qué punto se puede identificar algún tipo de variación lingüística en el corpus objeto de estudio. En cuanto a la segunda pregunta, se trata de averiguar, en el caso de que estas variaciones existan, hasta qué punto pueden tener su origen en el efecto de imprimación de la traducción automática.

Para llevar a cabo el estudio se ha empleado una metodología cuantitativa para explorar la diversidad y variedad léxica, la longitud del corpus y un conjunto de patrones léxicos. Así, en respuesta a la primera pregunta de investigación, se han podido identificar los siguientes cuatro indicios de variación lingüística en el corpus objeto de estudio: 1) el léxico de los textos traducidos automáticamente tiende a la simplificación y muestra menor diversidad léxica a lo largo del periodo de estudio; 2) aunque la densidad léxica de las palabras con contenido semántico (sustantivos, adjetivos, verbos y adverbios), muestra, *a priori*, una tendencia ascendente, los resultados parecen indicar que esta es compatible con el hecho de que la diversidad léxica de este mismo conjunto de palabras sea relativamente baja (es decir, que las palabras que se repiten sean sustantivos, adjetivos, verbos o adverbios); 3) la longitud entre el texto traducido y el original tiende a mantenerse similar; 4) la variación en los patrones léxicos muestra una ligera tendencia a descender, aunque, en este caso, los datos obtenidos no pueden considerarse realmente significativos, al indicar menos de 0,1 puntos en todos los casos. Los cuatro indicios parecen apuntar, entonces, a que existen ciertos rasgos de variación lingüística que afectan a los textos traducidos automáticamente.

Esto nos lleva a la segunda pregunta de investigación, para determinar si esta variación puede tener su origen en el efecto de imprimación de la traducción automática. En este sentido, los resultados obtenidos están en consonancia con estudios como el que presentan Vanmassenhove et al. (2021), quienes indican que se da una correlación entre la riqueza lingüística de los datos de entrenamiento de un sistema de traducción automática y su output en términos de diversidad morfológica y léxica. En su estudio, estos autores concluyen que las traducciones de los diferentes siste-

mas analizados muestran, de manera consistente, menor diversidad léxica que los datos empleados en el entrenamiento, es decir, que la capacidad de creatividad de un sistema de traducción automática se ve limitada (Vanmassenhof et al., 2021, p. 2204). Conviene tener en cuenta, en todo caso, que los sistemas de traducción automática que se emplean en la DGT no son estáticos, sino que están en continuo desarrollo para adaptarlos a las nuevas técnicas y reentrenarlos con datos adicionales. En este sentido, los resultados que se han presentado muestran la tendencia hasta el año 2020, pero no deben interpretarse como una predicción de cómo puede afectar el uso de los recientes desarrollos en inteligencia artificial aplicados a la traducción.

Un segundo aspecto que conviene mencionar es que, a diferencia de lo que ocurre con la traducción humana, la tecnología que opera en los sistemas de traducción automática carece de inteligencia o intencionalidad, de manera que para generar una traducción estos sistemas calculan probabilidades mediante algoritmos, pero no tienen la intencionalidad de modificar el texto traducido para adaptarlo a las características únicas de un determinado género textual ni a los criterios de aceptabilidad de un usuario potencial (Jiang y Niu, 2022). En todo caso, es necesario puntualizar que es posible que las características únicas de un género textual sean reproducidas exitosamente de modo probabilístico, pero esto no obedecerá a una intencionalidad o inteligencia. Si sucede, será por motivos estadísticos. Además, tal y como sostienen Vanmassenhove et al. (2021), la naturaleza inherentemente probabilística del algoritmo de los sistemas de traducción automática puede hacer que estos empleen de manera recurrente las palabras más frecuentes sin tener en cuenta las que se dan con menor frecuencia en los datos de entrenamiento, es decir, que se produce un efecto de sesgo algorítmico o sesgo estadístico. Todo esto podría dar como resultado una menor diversidad léxica de los textos traducidos automáticamente.

Por último, cabe señalar que la presentación de los resultados cuantitativos en el análisis de corpus que hemos llevado a cabo, si bien proporciona una visión amplia de las tendencias lingüísticas, tiene algunas limitaciones. Los datos numéricos ofrecen información valiosa sobre la frecuencia y distribución de ciertos fenómenos, pero no permiten captar matices semánticos, pragmáticos o contextuales que son esenciales para una comprensión profunda del fenómeno que observamos. El enfoque cuantitativo tiende a simplificar o generalizar realidades complejas, lo que puede llevar a interpretaciones incompletas o sesgadas. Por ello, resulta indispensable complementar el análisis que hemos presentado con una perspectiva cualitativa que permita examinar casos individuales, analizar el contexto de uso de las palabras o estructuras y explorar las variaciones lingüísticas que no se reflejan en las cifras. Solo mediante la combinación de ambos enfoques –cuantitativo y cualitativo– es posible obtener una comprensión integral y matizada de los fenómenos lingüísticos analizados, enriqueciendo así los resultados del estudio. Asimismo, sería necesario abordar el estudio del efecto de imprimación en el conjunto de lenguas de la UE y comprobar si existe alguna correlación con los datos que aquí se han presentado. De igual manera, sería interesante averiguar qué influencia puede tener el género textual en el efecto de imprimación de la traducción automática,

ya que cada género presenta convenciones lingüísticas, sintácticas y estilísticas específicas. Textos técnicos o especializados, por ejemplo, suelen seguir estructuras rígidas y terminología precisa, lo que facilita la uniformidad en las traducciones automáticas y refuerza la dependencia del sistema en patrones previos. Por otro lado, géneros más creativos o literarios, al requerir mayor flexibilidad interpretativa, pueden generar un efecto de imprimación más limitado, donde el traductor humano necesita realizar ajustes importantes. Así, el impacto de la imprimación podría variar en función de las demandas contextuales y estilísticas del cada género textual. Como vemos, el estudio del efecto de imprimación de la traducción automática y la exploración de sus características únicas puede ofrecer nuevas vías de investigación en los estudios de traducción basados en corpus y aportar una comprensión profunda de las características lingüísticas propias de las traducciones automáticas.

REFERENCIAS BIBLIOGRÁFICAS

Anthony, L. (2022). *TagAnt* (Version 2.0.5) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/softwar

Anthony, L. (2023). *AntConc* (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/softwar

Aranberri, N. (2020). Can translationese features help users select an MT system for post-editing? *Procesamiento del Lenguaje Natural*, *64*, 93-100. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/620

Avramidis, E., Burchardt, A., Hunsicker, S., Popovic, M., Tscherwinka, C., Vilar, D., y Uszkoreit, H. (2014). The taraxu¨ corpus of human-annotated machine translations. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2679–2682. http://www.lrec-conf.org/proceedings/lrec2014/pdf/401_Paper.pd

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. En M. Baker, G. Francis, y E. Tognini-Bonelli (Eds.), *Text and technology: in honour of John Sinclair* (pp. 233-252). John Benjamins.

Bangalore, S., Behrens, B., Carl, M., Ghankot, M., Heilmann, A., Nitzke, J., Schaeffer, M., y Sturm, A. (2016). Syntactic Variance and Priming Effects in Translation. En M. Carl, S. Bangalore, y M. Schaeffer (Eds.), *New Directions in Empirical Translation Process Research. New Frontiers in Translation Studies* (pp. 211-238). Springer.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., y Zampieri, M. (2016, agosto). *Findings of the 2016 conference on machine translation (wmt16). First conference on machine translation*, 131-198. https://aclanthology.org/W16-2301

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., HerbertVoss, S., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray,

S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., y Amodei, D. (2020). Language models are few-shot learners. *Advances in NeurIPS*, *33*, 1877–1901.

CELEX. (1994). Communication from the Commission to the Council and the European Parliament. *Final evaluation of the results of Eurotra: a specific programme concerning the preparation of the development of an operational Eurotra system for Machine Translation*. Document 51994DC0069. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:51994DC006

Cettolo, M., Niehues, J., Stuker, S., Bentivogli, L., Cattoni, R., y Marcello, F. (2015). The iwslt 2015 evaluation campaign. *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, 2–14. https://aclanthology.org/2015.iwslt-evaluation.1

Čulo, O., y Nitzke, J. (2016). Patterns of Terminological Variation in Post-editing and of Cognate Use in Machine Translation in Contrast to Human Translatio. *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, 106–114. https://aclanthology.org/W16-3401

Daems, J., De Clercq, O., y Macken, L. (2017). Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 89-103.

DataLitMT (2023). *Data Evaluation: Machine Translationese and Post-Editese*. https://github.com/ITMK/DataLitMT/blob/main/learning_resources/machine_translationese_post-editese/Data_Evaluation_Machine_Translationese_and_Post-Editese_Basic_Paper.pd Recuperado el 17 de julio de 2024.

DGT-Translation memory. (2024). https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memor Recuperado el 17 de julio de 2024.

EUR-Lex (1997). EUR-Lex - 91996E2286 – en. Written Question No. 2286/96 by Ben FAYOT to the Commission. Systran translation system developed by DG XIII - Position of the system development team. *Official Journal C 011 , 13/01/1997*. *https://eur-lex.europa.eu/legal-content/MT/TXT/?uri=CELEX:91996E00228*

EUR-Lex (2012). *Carta de los derechos fundamentales de la Unión Europea. (2012/C 326/02). https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:12012P/TX*

European Commission. (2008). *Translation Tools and Workflow*. European Commission Directorate-General for Translation Communication and Information Unit.

European Commission. (2016). *Translation Tools and Workflow*. European Commission Directorate-General for Translation Communication and Information Unit.

Farrel, M. (2018). Machine Translation Markers in Post-Edited Machine Translation Output. *Proceedings of the 40th Conference Translating and the Computer*, 50–59.

Foti, M. (2012, 29 y 30 de noviembre). *MT@EC: Working with translators*. [Paper presentation]. ASLIB - Translating and the Computer Conference, Londres, Reino Unido. https://aclanthology.org/2012.tc-1.4.pd

Freitag, M., Vilar, D., Grangier, D., Cherry, C., y Foster, D. (2022). A Natural Diet: Towards Improving Naturalness of Machine Translation Outpu. *Findings of the Association for Computational Linguistics: ACL*, 3340–3353. https://aclanthology.org/2022.findings-acl.263

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

Jian, Y. y Niu, J. (2022). A corpus-based search for machine translationese in terms of discourse coherence. *Across Languages and Cultures*, *23*(2), 148–166.

Jimenez-Crespo, M. A. (2023). "Translationese" (and "post-editese"?) no more: on importing fuzzy conceptual tools from Translation Studies in MT researc. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation,* 261–268. European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.25

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Linguistics and Phonetics Working Papers*, *53*, 61-79.

Kajzer-Wietrzny, M., y Ivaska, I. (2020). A multivariate approach to lexical diversity in constrained language. *Across Languages and Cultures*, *21*(2), 169–194.

Kyle, K. (2020). *Lexical-diversity 0.1.1* [Computer software]. https://pypi.org/project/lexical-diversity/#descriptio

Koehn, P., Hoang, H., Birch, A. Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., y Herbst, E. (2007). *Moses: Open Source Toolkit for Statistical Machine Translatio. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume Proceedings of the Demo and Poster Sessions*,177–180. https://aclanthology.org/P07-2045

Krüger, H., y van Rooy, B. (2016). Constrained language: a multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, *37*(1), 26–57.

Krüger, R. (2020). Explicitation in Neural Machine Translation. *Across Languages and Cultures*, *21*(2), 195-216 (2020). https://doi.org/10.1556/084.2020.0001

Lapshinova-Koltunski, E. (2013). VARTRA: A Comparable Corpus for the Analysis of Translation Variation. *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, *77*–86. https://aclanthology.org/W13-2510

Lapshinova-Koltunski, E. (2015). Variation in Translation: Evidence from Corpora. En C. Fantinuoli, y F. Zanettin (Eds.), *New Directions in Corpus-Based Translation Studies* (pp. 93–113). Translation and Multilingual Natural Language Processing 1. Language Science Press.

Maegaard, B. (1995). EUROTRA, History and Results. *Proceedings of the V MT Summitt. Luxemburgo*. https://aclanthology.org/1995.mtsummit-1.5.pd

Maegaard, B., y Perschke, S. (1991). An Introduction to the Eurotra Programme. En C. Copeland, J. Durand, S. Krauwer, y B. Maegaard (Eds.), *The Eurotra Linguistic Specifications. Studies in Machine Translation and Natural Language Processing*, vol. 1 (pp. 7-14). Office for Official Publications of the Commission of the European Community.

Martikainen, H., y Kuble, N. (2016). Ergonomie cognitive de la post-edition de traduction automatique: enjeux pour la qualite des traductions. *ILCEA Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie, 27*, 1-17. https://doi.org/10.4000/ilcea.386

McCarthy, P. M., y Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.38

Niu, J., y Jiang, Y. (2024). Does simplification hold true for machine translations? A corpus-based analysis of lexical diversity in text varieties across genres. *Humanities and Social Sciences Communications,* 11, 1-10. https://doi.org/10.1057/s41599-024-02986-

Richards, B. (2006). Type/Token Ratios: what do they really tell us? *Journal of Child Language, 14*, 201 – 209.

Rico Pérez, C. (2024). *Corpus UCM-EUROPA: estudio del efecto de imprimación de la traducción automática sobre un corpus de textos del español institucional* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1401718

Rothwel, A., Moorken, J., Fernández-Parr, M., Druga, J., y Austermueh, F. (2023). *Translation Tools and Technologies*. Routledge.

Salton, G. (1989). *Automatic text processing: The transformation, analysis and retrieval of information by computer*. Addison-Wesley Longman.

Sánchez Ramos, M. M., y Rico Pérez, C. (2020). *Traducción automática. Conceptos clave, procesos de evaluación y técnicas de posedición*. Comares.

Sánchez-Saus Laserna, M. (2022). ¿De qué hablamos cuando divulgamos sobre lingüística? Análisis de un corpus de textos divulgativos y aplicaciones al estudio terminológico de la semántica léxica. *ELUA. Estudios de Lingüística, 38*, 73-98. https://doi.org/10.14198/ELUA.2238

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.

Steinberger R., Eisele, A., Klocek, S., Pilos, S., y Schlüter, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Language. *Proceedings of the 8th international conference on Language Resources and Evaluation* (LREC'2012). https://aclanthology.org/L12-1481

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., y Gilbro, S. (2014). An overview of the European Union's highly multilingual parallel corpor . *Language Resources and Evaluation Journal (LRE)*, 679-707. https://doi.org/10.1007/s10579-014-9277-

Teich, E. (2003). *Cross-linguistic variation in system and text: a methodology for the investigation of translations and comparable texts*. Mouton de Gruyter.

Tirkkonen-Condit, S. (2004). Unique items: over- or under-represented in translated language? En A. Mauranen, y P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 177-184). John Benjamins.

Toma, P. (1976). SYSTRA. *Foreign Broadcast Information Service Seminar on Machine Translation*, 40 –45. https://aclanthology.org/1976.earlymt-1.11

Toral, A., y Sánchez-Cartagena, V.M. (2017). A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Volume 1, 1063–1073. https://aclanthology.org/E17-1000

Toral, A. (2019). Post-editese: an Exacerbated Translationes. *Proceedings of Machine Translation Summit XVII: Research Track,* 273–281. https://aclanthology.org/W19-6627

Translation Centre. (2019). *Consolidated Activity Report of the Translation Centre 2018*. Translation Centre for the Bodies of the European Union. https://op.europa.eu/s/zFA

Translation Centre. (2023). *Consolidated Activity Report of the Translation Centre* 2022, Translation Centre. https://data.europa.eu/doi/10.2817/76991

Vanmassenhove, E., Shterionov, D., y Gwilliam, M. (2021). Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translatio. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2203–2213. https://aclanthology.org/2021.eacl-main.188.pd

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser. L., y Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. https://doi.org/10.48550/arXiv.1706.0376

Volkart, L., y Bouillon, P. (2022). Studying Post-Editese in a Professional Context: A Pilot Study. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 71–79. https://aclanthology.org/2022.eamt-1.10