ISSN: 2171-6633

LA CAPACIDAD DE APRENDIZAJE DE LA TA NEURONAL: UN ESTUDIO LONGITUDINAL

ASTRID SCHMIDHOFER
Universität Innsbruck
Astrid.Schmidhofer@uibk.ac.at

Fecha de recepción: 18.12.2021
Fecha de revisión: 08.01.2022
Fecha de aceptación: 23.01.2022

Resumen: En este trabajo se presentarán los resultados de un estudio longitudinal de dos sistemas abiertos de TA neuronal. Se compararon las traducciones, de siete textos de características diferentes, que se habían obtenido en tres momentos distintos a lo largo de un año en la combinación español-alemán.

Palabras clave: traducción automática, traducción neuronal, traducción español/alemán, GoogleTranslate, DeepL.

The Learning Capacity of Neuronal MT: A Longitudinal Study

Abstract: This paper presents the results of a long-term study of Spanish to German translations provided by two open-access NMT systems. The study was carried out with seven different texts that were introduced into the systems at various moments throughout a year.

Key words: machine translation, NMT, Spanish/German translation, GoogleTranslate, DeepL

1. Introducción

Cuando hace unos años aparecieron los primeros sistemas de TA neuronal, las empresas desarrolladoras afirmaban que una de las mayores ventajas que ofrecían estos frente a sistemas anteriores consiste en que son capaces de ir mejorando la calidad de sus traducciones gracias a la tecnología del aprendizaje profundo (*deep learning*). Emulando los procesos del cerebro humano y nutriéndose de grandes corpus de datos y correcciones por parte de revisores humanos, estos sistemas conseguirían ir subiendo la calidad de sus traducciones e ir acercándose poco a poco a la calidad de los productos elaborados por traductores profesionales humanos (cf. Wu/ Schuster/Chen/Le/ Norou 2016: 2 y Hassan et al. 2018).

Compartimos la impresión de que los primeros sistemas neuronales accesibles de forma gratuita (nos referimos particularmente a GoogleTranslate y DeepL) mostraban, a primera vista, una calidad superior a sistemas abiertos de generaciones anteriores. Su lanzamiento generó muchas expectativas, incluso entusiasmo, por un lado (cf. Le/Schuster 2016; Plass-Fleßenkämper 2017), y predicciones apocalípticas para la profesión del traductor, por otro (cf. Baureithel 2019; Elezaj 2019; Marr 2018). Es innegable que los sistemas neuronales están teniendo un impacto importante en muchos ámbitos, que abarcan desde la industria de la traducción[1] pasando por programas universitarios de formación de traductores hasta la forma en que los turistas interactúan con los oriundos en países cuya lengua no dominan.

En los últimos años se ha realizado un amplio número de estudios acerca de la calidad de la TA neuronal de los que presentaremos una selección en el epígrafe 2.2. Sin embargo, apenas existen estudios longitudinales que corroboren o cuestionen la afirmación expuesta más arriba de que los productos proporcionados por sistemas neuronales mejoran su calidad con el tiempo. La falta de estudios en este ámbito probablemente se debe, entre otras razones, a que los sistemas neuronales de acceso abierto solo llevan disponibles entre dos y tres años.

1 Sirva como ejemplo la conferencia con 1000 participantes organizada por la BDÜ en otoño de 2019 en Bonn cuyo título *Neue Wege im digitalen Zeitalter* y el gran número de comunicaciones y talleres relacionados con la TA reflejan el enorme interés de la industria por la traducción automática: https://www.uebersetzen-in-die-zukunft.de/ (29.03.2020)

Para empezar a aportar datos en esta área, hemos llevado a cabo un estudio longitudinal en la combinación español-alemán con el objetivo de analizar los cambios que se observan en los productos proporcionados por los sistemas abiertos GoogleTranslate y DeepL a lo largo de un año. En este trabajo presentaremos los resultados globales de este estudio.

Nos parece importante destacar que este estudio se centra únicamente en el producto final presentado al usuario, que puede ser un traductor profesional o cualquier otra persona que desee obtener una traducción, y la evolución de dicho producto a lo largo del período descrito. No se trata, por lo tanto, de un estudio sobre la calidad de la TA frente a la traducción humana profesional ni sobre la arquitectura de los sistemas utilizados. Por este motivo, no profundizaremos en cuestiones técnicas y remitimos al lector interesado a Koehn (2020).

2 Estado de la cuestión

2.1 La TA neuronal: algunas observaciones

Los sistemas de TA neuronal están basados en redes neuronales artificiales profundas que se componen de varias capas muy conectadas que, a su vez, se componen de una serie de neuronas que procesan un *input* para generar un *output* activando funciones que previamente se han asignado a dichas neuronas. Dentro de la red, las palabras se representan con vectores únicos, llamados *word embeddings* en inglés. La generación del *output* se basa en la probabilidad percibida por el sistema con la que determinadas palabras aparecen juntas. Para una descripción más detallada, cf. Sánchez Ramos/Rico Pérez (2020: 21-26) y Casacuberta/Peris (2017).

Los dos sistemas que se han empleado en este análisis, GoogleTranslate y DeepL, utilizan arquitecturas diferentes. La primera versión de GoogleTranslate, que se publicó en 2016, se basaba en una arquitectura codificador-decodificador empleando dos redes neuronales recurrentes (RNN por sus siglas en inglés). Disponía asimismo de un mecanismo de atención mediante el cual se determina la importancia de cada palabra de una frase inicial mejorando así las predicciones de probabilidad. (cf. Wu/Schuster/Chen/Le/Norou 2016).

Los desarrollos más recientes se centran en los modelos *transformer* con los que se espera solucionar problemas persistentes en la TA automática con RNN como la traducción de frases largas o las referencias a elementos

alejados. Al igual que las RNN, esta tecnología utiliza un codificador y un decodificador, pero procesa todo el *input* al mismo tiempo y no por unidades individuales. Mediante un mecanismo de atención se analiza la relación de las palabras dentro de una misma frase. Se trata de "a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output" (Vaswani et al. 2017:2). Vaswani et al. (2017: 10) informan que su modelo "can be trained significantly faster than architectures based on recurrent or convolutional layers" y alcanzó puntuaciones BLEU excepcionales en las combinaciones inglés-alemán e inglés-francés. Según la información publicada por Google en su AI blog en junio de 2020, la última versión de GoogleTranslate emplea un codificador del tipo *transformer* y un decodificador RNN debido a los siguientes motivos: "Transformer models have been demonstrated to be generally more effective at machine translation than RNN models, but our work suggested that most of these quality gains were from the transformer encoder, and that the transformer decoder was not significantly better than the RNN decoder."(Caswell/Liang 2020).

La arquitectura que se esconde detrás de DeepL es un secreto bien guardado por la empresa desarrolladora. Lo poco que se sabe es que se emplean redes convolucionales neuronales (CNN por sus siglas en inglés), que se utilizan sobre todo en el reconocimiento de imagen, la tecnología *Beam Search* y un mecanismo de atención.

2.2. Estudios previos sobre la calidad de la TA neuronal

Existen, a día de hoy, ya muchos estudios acerca de la calidad de la TA neuronal, pues la evaluación de la calidad del *output* de la TA neuronal es de sumo interés tanto para la industria de la traducción como para la traductología. Para ello, se han desarrollado diferentes modelos de evaluación manual como el modelo SAEJ2450, MQM o el modelo TAUS y métricas para la evaluación automática como BLEU, METEOR y TER (cf. Sánchez Ramos/Rico Pérez 2020: 34-49). Son especialmente frecuentes los estudios que comparan la calidad del producto de sistemas de TA neuronal con la de otros tipos de traducción automática, especialmente la traducción estadística o la traducción estadística basada en frases (PBSMT por sus siglas en inglés). Los resultados obtenidos difieren considerablemente. El estudio

de Castilho/Moorkens/Gaspari/Calixto/Tinsley/Way (2017) concluye que la TA neuronal obtuvo muy buenos resultados en la evaluación automática, sin embargo, "the PBSMT system still produces better translation when assessed both via automatic and human evaluation metrics" (2017: 117). Casi contrarios con los resultados del estudio de Shterionow/Nagle/Casanellas/Superbo/O'Dowd (2018), quienes compararon la evaluación automática y la evaluación humana de sistemas neuronales y sistemas estadísticos PBMST y llegan a la siguiente conclusión: "While the quality evaluation scores indicated that the PBSMT engines perform better, the human reviewers show the opposite results, i.e., that NMT outperforms PBSMT. The human reviewers, all native speakers of the evaluated language pairs, ranked the quality of the NMT engines higher than that of PBSMT in all cases". En la prueba de productividad realizada por los investigadores se mostró que la mayoría de los traductores son más productivos en la posedición de un texto realizado por un sistema de TA neuronal comparado con la posedición de un texto de un sistema PBSMT y la traducción humana (Shterionow/Nagle/Casanellas/Superbo/O'Dowd, 2018: 233). Este último resultado contrasta, a su vez, con las conclusiones del estudio de López-Pereira (2019) sobre la productividad de la TA neuronal frente a la estadística, pues este revela que aunque los traductores participantes opinaron a priori que la posedición de un texto proporcionado por un motor de TA neuronal sería más productiva, se observó que a pesar de que la distancia de edición es menor, los participantes tardaron más en poseditar el texto propuesto por el motor neuronal.

Asimismo, se han realizado diferentes estudios que comparan la TA neuronal con la traducción humana. Queremos destacar, en este contexto, el estudio de Hassan et al. 2018, quienes afirman, para la traducción de noticias del chino al inglés, que "our latest neural machine translation system has reached a new state-of-the-art, and that the translation quality is at human parity when compared to professional human translations" (2018: 1) y la réplica de Läubli/Castilho/Neubig/Sennrich/Shen/Toral (2020:653), quienes muestran que „the finding of human-machine parity was owed to weaknesses in the evaluation design" y que "professional human translations contained significantly fewer errors".

Igualmente cabe mencionar que también se han llevado a cabo estudios enfocados a aspectos lingüísticos concretos como el de Chung (2018) sobre nombres propios y números y el de Mair/Schmidhofer (2019) sobre la traducción de expresiones relacionadas con el género gramatical. No hemos

podido encontrar, sin embargo, estudios longitudinales que evalúen la evolución de la calidad de productos de sistemas de TA neuronal a lo largo de un período extenso.

3 Estudio

3.1 Objetivo

El objetivo principal del presente estudio consistía en estudiar la evolución de dos sistemas abiertos de TA neuronal a lo largo de un año mediante el análisis de las traducciones proporcionadas por los sistemas en tres momentos diferentes a lo largo de este tiempo. Para tal fin, se comparó el *output* a principios, a mitad y al final de este período. Con este estudio se espera poder aportar datos para valorar si la calidad de los productos ofrecidos por estos sistemas mejora a lo largo del tiempo. La hipótesis principal que queda por comprobar es, por lo tanto, que la calidad de las traducciones mejora a lo largo de un año.

El material obtenido y el posterior análisis permiten, sin embargo, responder a toda una serie de preguntas que se tratarán en los diferentes subapartados del epígrafe 3.4. Estas son:

1. ¿Cómo se presenta la evolución de los textos en general si se categorizan los cambios según cambios a mejor (M), cambios neutrales (N) y cambios a peor (P)? (ep. 3.4.1)
2. ¿Qué resultados se obtienen de una comparación de los cambios entre las dos primeras versiones (*change 1*) y las dos segundas (*change 2*)? (ep. 3.4.2)
3. ¿En qué se diferencian los datos entre los dos sistemas empleados, DeepL y GoogleTranslate? ¿Cuál de los dos presenta más cambios? ¿Qué índices (mejor, neutral, peor) obtiene cada uno de ellos? (ep. 3.4.3)
4. ¿Cuántos cambios se observan en las diferentes categorías (contenido, léxico, morfosintaxis, estilo, ortografía, cf. la categorización más abajo) y cuál es la evolución de ellas?
5. ¿Existen diferencias entre los textos con respecto al número y la evolución de los cambios?

3.2 Material

Para este estudio se seleccionaron siete textos de una extensión de entre 275 y 381 palabras cada uno y con diferentes características. La muestra comprende textos puramente informativos (textos 1, 2 y 5), textos de carácter informativo-apelativo (textos 4 y 6), un texto expresivo (texto 7) y un texto de características mixtas (texto 3). Asimismo, tres textos contienen expresiones y nombres muy ligados a la cultura de origen (2, 4, 5), uno contiene terminología y nombres de política internacional (3), mientras que los demás no contienen expresiones asociadas a lugares determinados (1, 6, 7). En la tabla 1 se recoge información detallada acerca de los textos utilizados. Todos los textos seleccionados fueron escritos originalmente en español y provienen de España.

Texto	**Tipo de texto**	**Tema**
Texto 1	Folleto informativo	Tabaco y salud
Texto 2	Noticia de prensa sobre política española	Independentismo catalán
Texto 3	Editorial de prensa	Brexit
Texto 4	Folleto turístico	Pedraza en Segovia
Texto 5	Texto administrativo	Solicitud virtual
Texto 6	Folleto de publicidad	Préstamos personales
Texto 7	Texto literario (relato corto)	La iniciación a la literatura

Tabla 1: Textos seleccionados

3.3 Metodología

El procedimiento metodológico empleado puede subdividirse en cuatro fases: la selección de los textos, la traducción de los textos, la extracción de los datos y el análisis de los datos.

La selección de los textos se realizó a finales de diciembre de 2018. Los textos se eligieron según las características de la tabla y, en algunos casos, se recortaron para ajustarlos a la extensión deseada. Sin embargo, en todo momento, se procuró obtener textos coherentes.

Los textos se introdujeron en los sistemas elegidos (DeepL y GoogleTranslate) en tres momentos a lo largo de un año: a finales de diciembre de 2018, a principios de julio de 2019 y a finales de diciembre de 2019 recogiendo todas las versiones en un archivo de Excel.

La extracción de los datos se realizó de forma manual durante el mes de febrero de 2020. En total, se extrajeron 1054 entradas que corresponden siempre a cambios observados entre dos versiones consecutivas, es decir, entre la versión 1 y 2 o entre la versión 2 y 3. Las dificultades relativas a la extracción y la conformación de categorías se expondrán más abajo.

Finalmente se realizó un análisis estadístico cuyos resultados se presentan en el apartado 3.4. El análisis de datos se llevó a cabo con el programa SPSS 25. El nivel de significación se estableció en $p < .05$. Para las pruebas de hipótesis se utilizaron tres procedimientos:

- la prueba χ^2 bidimensional que es aplicable a datos de escala nominal;
- la prueba de la U de Mann-Whitney para datos de escala ordinal y para comparar dos grupos independientes;
- la prueba de Kruskal-Wallis que permite comparar datos de varios grupos de datos de escala ordinal.

3.3.1 Cuestiones relativas a la extracción de datos y la categorización de los cambios

Para extraer los cambios entre las diferentes versiones se compararon primero la versión 1 y la versión 2 (*change* 1) y posteriormente la versión 2 y la versión 3 (*change* 2) de cada texto. Se anotaron los cambios observados en una tabla Excel, se categorizaron según las categorías que se indican a continuación y se les asignó una valoración (mejor, neutral, peor). La comparación se realizó de forma automática mediante un programa informático, sin embargo, la extracción de los cambios y su inserción en la tabla Excel se tuvieron que realizar de forma manual puesto que la segmentación de las oraciones, la identificación de los cambios, la asignación de la categoría y la valoración es un trabajo altamente especializado para el que se requiere la experiencia de una persona habituada a la evaluación de traducciones.

Con respecto a la identificación de los cambios, cabe destacar que cualquier tipo de modificación marcada en la comparación automática y comprobada de forma manual se identificó como cambio y se recogió en la tabla. En el caso de dos cambios completamente idénticos en un mismo texto (por ejemplo dos veces el mismo error ortográfico, dos veces el mismo cambio léxico), estos solo se contabilizaron una vez.

En relación a las categorías, tras mucha deliberación decidimos aplicar un sistema sencillo con cinco categorías: contenido, léxico, morfosintaxis, estilo y ortografía[2]. La decisión de optar por una clasificación de pocas categorías radica en que facilita tanto la asignación de los cambios a las diferentes categorías como el posterior análisis estadístico y la presentación.

Las diferentes categorías podrían describirse de la siguiente manera:

Tipo	**Explicación**	**Unidad que abarca**
Contenido (C)	Se percibe un cambio en el mensaje del texto.	oraciones completas o partes de oraciones que configuran una unidad de sentido más allá de una sola unidad léxica, ocasionalmente unidades léxicas aisladas cuyo cambio conlleva un impacto importante en el mensaje del texto
Léxico (L)	Se han empleado dos unidades léxicas distintas, pero no se aprecia un cambio sustancial en el mensaje del texto.	unidades léxicas simples y compuestas
Morfosintaxis (G)	Se aprecia un cambio dentro de una estructura morfosintáctica sin que este conlleve una modificación sustancial del mensaje.	sintagmas nominales, sintagmas verbales, oraciones completas a nivel sintáctico
Estilo (S)	Se aprecia un cambio en la forma de expresar una idea.	oraciones completas o partes de oraciones (por ejemplo estilo nominal/verbal)
Ortografía (O)	Se observa un cambio a nivel ortográfico o de puntuación.	unidades léxicas, signos de puntuación, símbolos

Tabla 2: Sistema de categorías de cambios

2 Barajamos la posibilidad de aplicar o adaptar sistemas de evaluación manual como, por ejemplo, el sistema MQM que garantizan una mayor comparabilidad entre estudios. Sin embargo, estos sistemas nos parecían más indicados para la anotación de errores y no tanto para la evaluación de cambios.

Como algunas categorías atienden a la forma y otras al mensaje, en muchas ocasiones se aprecian diferentes cambios en una misma oración o un mismo sintagma. También existe un solapamiento ocasional entre las categorías de contenido y léxico por un lado y contenido y morfosintaxis por otro. En el primer caso, en algunas ocasiones era difícil decidir en qué medida la sustitución de una unidad léxica por otra afecta sustancialmente al sentido; en el segundo, en qué medida un cambio en la estructura morfosintáctica puede modificar el mensaje.

El reto más grande consistía, sin duda alguna, en la asignación de una valoración al cambio anotado. Decidimos establecer un sistema simple de tres categorías: cambio a mejor (M), cambio neutral (N) y cambio a peor (P). En las categorías formales (G y O), el cambio podía cotejarse con la norma establecida por lo cual la valoración resultó bastante clara y sencilla. No obstante, en las demás categorías surgieron algunas dificultades en el proceso de valoración. En la categoría de contenido, por ejemplo, a veces tuvieron que contrastarse dos fragmentos carentes de sentido; si estos cambios no condujeron a una mejor comprensión, se les asignó la valoración N. Las categorías que más retos plantearon fueron las categorías L y S. La valoración de la adecuación de una unidad léxica o de un recurso estilístico no siempre es objetivable. Para minimizar el cariz subjetivo, se estudiaron textos paralelos en la lengua meta y, en casos dudosos, se consultaron colegas de profesión. Como norma general, si el cambio no era objetivamente a mejor o peor, se asignó la categoría N.

3.4 Resultados y discusión

3.4.1. Observaciones generales

El número total de cambios observados asciende a 1054 para todos los textos que en conjunto comprenden 2252 palabras. Sin embargo, no debe olvidarse que esta cifra de cambios comprende tanto los dos sistemas utilizados (DeepL y GoogleTranslate) como las dos comparaciones realizadas en dos períodos consecutivos (*change* 1: versión 1 > versión 2 y *change* 2: versión 2 > versión 3). Aun así, el número de datos superó ampliamente nuestras expectativas. El alto número de ítems obtenido es indicativo de que los dos sistemas sufren cambios considerables a lo largo de un tiempo relativamente corto.

Asimismo, cabe destacar que en ocasiones se observan los mismos cambios (o cambios muy similares) en los dos sistemas. También es llamativo que en algunos elementos se observa un cambio en el período *change* 1 (versión 1 > versión 2) que, sin embargo, es revertido en el período *change* 2, por lo que la versión 3 coincide con la versión 1 en estos ítems. Se trata, sin embargo, de meras observaciones generales para cuya comprobación haría falta un estudio aún más detallado de los datos obtenidos.

Los datos generales que se han obtenido en el análisis son los siguientes: a 375 de los cambios anotados se asignó la valoración M (cambio a mejor), a 386 la valoración N (cambio neutral) y a los restantes 293 la valoración P (cambio a peor). Prevalecen, por lo tanto, los cambios neutrales y a mejor, lo que sí confirma que los sistemas analizados mejoran a lo largo del tiempo.

Tipo de cambio	**Número total**	**Porcentaje**
M	375/1054	35,58 %
N	386/1054	36,62 %
P	293/1054	27, 8%

Tabla 3: Valoración global de los cambios observados

3.4.2. Evolución de las traducciones

El análisis de la evolución de las traducciones se realizó mediante un análisis de los cambios entre dos versiones consecutivas, es decir, se compararon las versiones 1 y 2 cuyos cambios se designaron *change* 1, y después las versiones 2 y 3, cuyos cambios se designaron *change* 2. Se optó por este procedimiento dado que no se trató de una valoración independiente de los tres textos, sino de analizar la evolución de los mismos a lo largo de un período mediante la comparación de los cambios observables en sus respectivas traducciones. El procedimiento elegido permite cuantificar los cambios a lo largo del tiempo de medición y a través de las tres versiones.

La comparación entre las versiones 1 y 2 dio como resultado un total de 453 cambios, entre los cuales prevalecen los cambios neutrales y aquellos a peor. En la comparación entre las versiones 2 y 3, el número total de cambios obtenidos es mayor y asciende a 601. En esta comparación, destacan claramente los cambios a mejor.

Datos generales

***Change* 1 (versiones 1>2)**			***Change* 2 (versiones 2>3)**		
Tipo de cambio	Número total	Porcentaje	Tipo de cambio	Número total	Porcentaje
M	128/453	28,26 %	M	247/601	41,1 %
N	163/453	35,98 %	N	223/601	37,1 %
P	162/453	35,76 %	P	131/601	21,8 %

Tabla 4: Valoración de los cambios por períodos

Para explorar estas diferencias con ayuda de un procedimiento estadístico, se adoptó como hipótesis nula que no existen diferencias significativas entre *change* 1 y *change* 2 y como hipótesis alternativa que tales diferencias sí se dan. Para responder a esta cuestión se utilizó la prueba de la U de Mann-Whitne. El resultado de esta prueba es significativo con un valor $p < .001$, lo que confirma que existen diferencias significativas entre *change* 1 y *change* 2. Los resultados porcentuales se muestran en el gráfico 1.

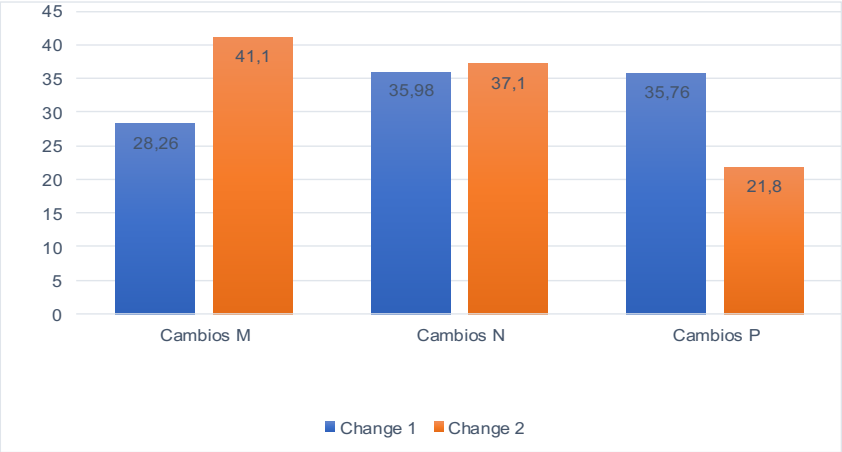

Gráfico 1: Comparación entre los dos períodos de medición (porcentajes)

Estos datos parecen indicar que los sistemas de TA neuronal utilizados muestran un alto grado de fluctuación a lo largo del tiempo. Es especialmente llamativo el alto índice de cambios a peor en el primer período que se revierte a un porcentaje importante de cambios a mejor en el segundo.

3.4.3. Comparación entre los dos sistemas

Si comparamos los datos de los dos sistemas empleados (DeepL y GoogleTranslate), obtenemos los siguientes datos generales: el número total de cambios es ostensiblemente menor en DeepL (408) que en GoogleTranslate (646). En los diferentes subgrupos, se puede observar un índice ligeramente más alto en los cambios a mejor (36,76 % vs. 34, 83 %) y un índice visiblemente más bajo en los cambios a peor (24,75 % vs. 29, 72 %) a favor de DeepL.

DeepL			**GoogleTranslate**		
Tipo de cambio	Número total	Porcentaje	Tipo de cambio	Número total	Porcentaje
M	150/408	36,76 %	M	225/646	34,83 %
N	157/408	38,48 %	N	229/646	35,45 %
P	101/408	24,75 %	P	192/646	29,72 %

Tabla 5: Comparación en la valoración de los cambios de DeepL y GoogleTranslate

Para averiguar si existen diferencias significativas entre los dos sistemas relativas a la valoración de los cambios, se establece como hipótesis nula la no-existencia de tales cambios y como hipótesis alternativa su existencia. En el análisis se utilizó, de nuevo, la prueba de la U de Mann-Whitney. Como resultado se obtuvo una diferencia no significativa de $p = .18$, lo que obliga a mantener la hipótesis nula puesto que se halla por encima del límite de p = .05. No se aprecian, por lo tanto, diferencias significativas entre los dos grupos. El gráfico 2 muestra una contrastación de los valores porcentuales.

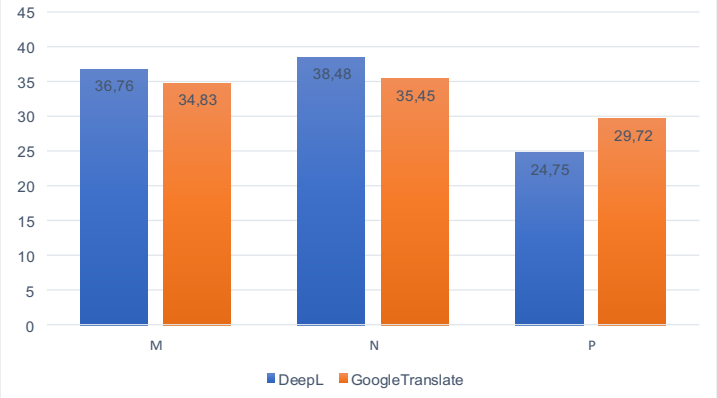

Gráfico 2: Comparación entre DeepL y GoogleTranslate (porcentajes)

3.4.4. Comparación de las diferentes categorías

Tal como se explicó en el apartado 3.3.1, los cambios se asignaron a las siguientes cinco categorías: contenido (C), léxico (L), morfosintaxis (G), estilo (S), ortografía (O). La categoría con más cambios es la de léxico, con 413 ítems, seguida por la de morfosintaxis, con 250, y las de contenido y estilo, con 178 y 176 respectivamente. La categoría con menos ítems es la de ortografía, con tan solo 37. En la tabla 6 se pueden ver los cambios por categoría desglosados según los parámetros de valoración (M/N/P).

Categoría	**Número total**	**Porcentaje**	**Valoración**	**Número total**	**Porcentaje**
C	178/1054	16,89 %	M	91	51,12 %
			N	29	16,29 %
			P	58	32,58 %
L	413/1054	39,18 %	M	127	30,75 %
			N	194	46,97 %
			P	92	22,28 %
G	250/1054	23,72 %	M	104	41,6 %
			N	57	22,8 %
			P	89	35,6 %
S	176/1054	16,7 %	M	40	22,73 %
			N	98	55,68 %
			P	38	21,59 %
O	37/1054	3,51 %	M	13	35,14 %
			N	8	21,62 %
			P	16	43,24 %

Tabla 6: Valoración de los cambios por categoría

La categoría con más cambios es con diferencia la de léxico. Esto se debe a que la mayoría de los cambios se observan en la sustitución de un elemento de naturaleza semántica por otro. La categoría con menos ítems es, también con mucha diferencia, la de ortografía, que es una categoría bastante estable debido al alto grado de fijación de la norma ortográfica.

El alto índice de cambios neutrales en las categorías L y S se debe a que se trata de las categorías más "subjetivas" en las que la valoración depende más de la percepción del evaluador que en otras. Intentamos categorizar como M y P solo aquellos ítems en los que el cambio a mejor/peor era objetivable. En cambio, las categorías con una normativa más clara y rígida, G y O, comprenden menos cambios neutrales. Para analizar si existen diferencias significativas entre las categorías con respecto a la valoración, se

utilizó la prueba de Kruskal-Wallis con la que se obtuvo un valor de $p = .14$. No existen, por lo tanto, diferencias significativas entre las diversas categorías en relación a la valoración.

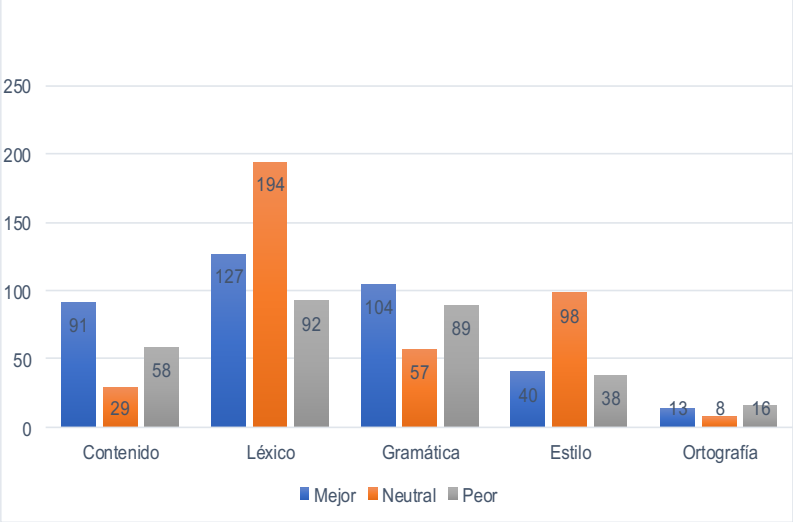

Gráfico 3: Valoración de los cambios por categorías (números absolutos)

Asimismo se analizaron las categorías en relación a los cambios de la versión 1 a la versión 2 (*change* 1) y de la versión 2 a la versión 3 (*change* 2).

Categoría	**Período**	**Número total**	**Porcentaje**
C	*Change* 1	77/453	17 %
	Change 2	101/601	16,81 %
L	*Change* 1	167/453	36,87 %
	Change 2	246/601	40,93 %
G	*Change* 1	115/453	25,39 %
	Change 2	135/601	22,46 %
S	*Change* 1	73/453	16,11 %
	Change 2	103/601	17,14 %
O	*Change* 1	21/453	4,64 %
	Change 2	16/601	2,66 %

Tabla 7: Evolución de las categorías a lo largo del tiempo de medición

A primera vista, no se aprecian diferencias destacables entre los dos períodos. Para comprobar esta impresión mediante un análisis más preciso, se aplicó, de nuevo, un test χ^2 bidimensional que da un resultado no significativo de p = .28 con lo que se confirma que las categorías aparecen en ambos períodos con una frecuencia comparable.

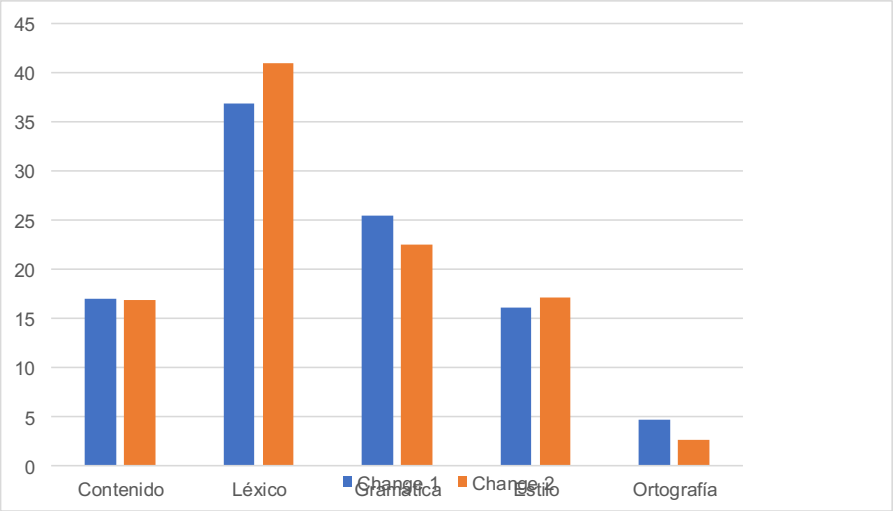

Gráfico 4: Categorías en los dos períodos de evaluación (porcentajes)

Por último, comparamos la aparición de las diferentes categorías en los dos sistemas utilizados.

Categoría	Sistema	Número total	Porcentaje
C	DeepL	77/408	18,87 %
	GoogleT	101/646	15,63 %
L	DeepL	161/408	39,46 %
	GoogleT	252/646	39 %
G	DeepL	71/408	17,4 %
	GoogleT	179/646	27,71 %
S	DeepL	89/408	21,81 %
	GoogleT	87/646	13,47 %
O	DeepL	10/408	2,45 %
	GoogleT	27/646	4,18 %

Tabla 8: Reparto de los ítems de las categorías según los sistemas

Para analizar el reparto de las categoría según los dos sistemas utilizados se aplicó, de nuevo, un test χ^2 bidimensional que da un resultado significativo de $p < .001$. Se aprecian, por lo tanto, diferencias significativas entre los dos sistemas en relación a la frecuencia de las categorías. Estas son particularmente visibles en las categorías G y S. Mientras en GoogleTranslate, los cambios de índole morfosintáctica (G) son notablemente más frecuentes que en DeepL, ocurre justo lo contrario en la categoría de estilo (S).

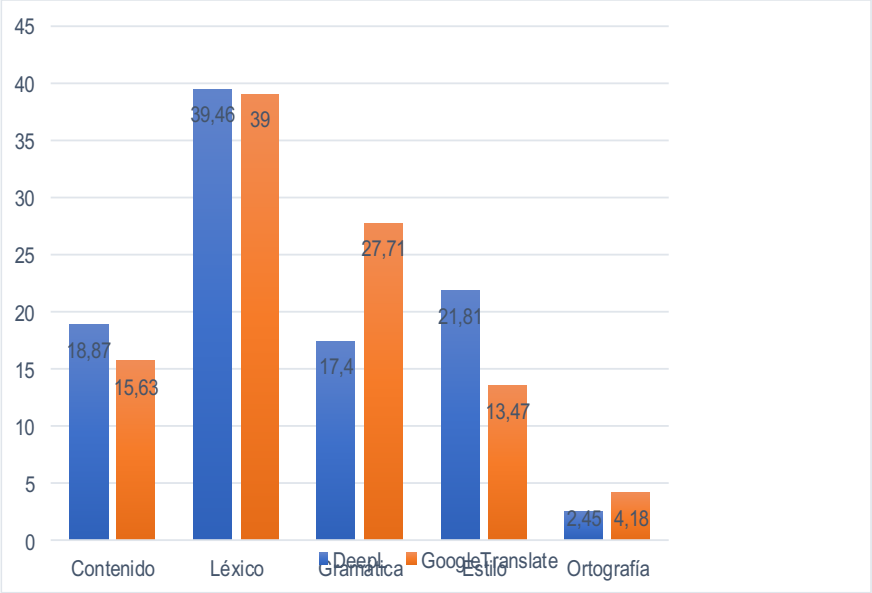

Gráfico 5: Relación entre las categorías y los sistemas (porcentajes)

3.4.5. Comparación entre los textos

Por último, se realizó una comparación entre los textos que se habían elegido para este estudio. En la tabla 10, se ven desglosados los cambios y la valoración de estos por cada texto.

Texto	Número de palabras y porcentaje del total	Número total de cambios	Porcentaje sobre el total	Cambios a mejor (M)	Cambios neutrales (N)	Cambios a peor (P)
1	329 (14,61 %)	89/1054	8,44 %	29 (32,58 %)	42 (47,19 %)	18 (20,22 %)
2	333 (14,79 %)	186/1054	17,65 %	64 (34,41 %)	66 (35,48 %)	56 (30,11 %)
3	381 (16,92 %)	203/1054	19,26 %	74 (36,45 %)	74 (36,45 %)	55 (27, 09 %)
4	291 (12,92 %)	130/1054	12,33 %	41 (31,54 %)	52 (40 %)	37 (28,46 %)
5	300 (13,32 %)	111/1054	10,53 %	33 (29,73 %)	46 (41,44 %)	32 (28,82 %)
6	275 (12,21 %)	103/1054	9,77 %	38 (36,89 %)	39 (37,86 %)	26 (25,24 %)
7	343 (15,23 %)	232/1054	22,01 %	96 (41,38 %)	67 (28,88 %)	69 (29,74 %)

Tabla 9: Desglose de los cambios por texto

Los textos que muestran el menor número de cambios y, por lo tanto, los más estables son los textos 1, 5 y 6. Este resultado no sorprende en el caso del texto 1 que es puramente informativo y versa sobre una temática global. El mismo carácter informativo se observa en el texto 5 aunque este contiene varias expresiones propias de un sistema administrativo determinado (España), lo que suele plantear problemas a los traductores (automáticos). Por último, el texto 6 también versa sobre una temática global. A pesar de que tiene también carácter apelativo, el fuerte componente informativo y su contenido universal parecen haber contribuido a que los cambios han sido relativamente pocos a lo largo del tiempo de medición.

En el otro extremo encontramos los textos 2, 3, 7. Los tres textos y, en particular el texto 7, muestran una mayor complejidad sintáctica frente a los otros textos y una riqueza léxica que en ocasiones adquiere un carácter metafórico y/o figurado. Estas características suponen mayores problemas de traducción, lo que podría explicar la mayor fluctuación observada en estos tres textos.

Para analizar si existe una dependencia relevante a nivel estadístico entre los diferentes textos y la valoración de los cambios, se aplicó, de nuevo,

la prueba de Kruskal-Wallis, que dio como resultado p = .82, lo que significa que no se aprecian diferencias significativas.

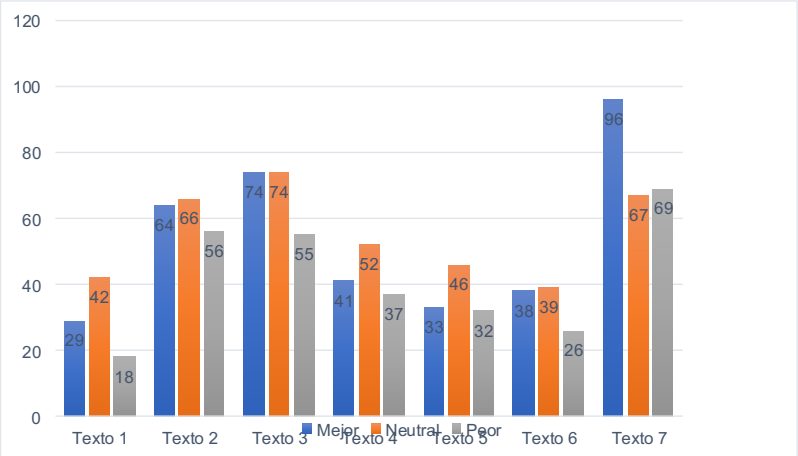

Gráfico 6: Valoración de los cambios por texto (números absolutos)

3.4.6. Impacto

Por último, intentamos establecer el impacto de los cambios para cada texto mediante un valor numérico. Para ello, asignamos valores descendientes por orden de importancia a las categorías, a saber: C =5, L = 4, G = 3, S = 2, O = 1. Asimismo, asignamos un punto positivo (+1) por cada cambio a mejor y un punto negativo (-1) por cada cambio a peor. Los cambios neutrales no recibieron puntos. El resultado del cálculo de impacto se encuentra desglosado en la tabla 10.

Texto	**Categoría**	**Cambio a mejor (M)**	**Cambio neutral (N)**	**Cambio a peor (P)**	**Puntuación**	**Impacto por texto**
1	C	3	0	0	15	46
1	L	15	18	5	40	
1	G	3	12	6	-9	
1	S	6	11	7	-2	
1	O	2	1	0	2	
2	C	14	7	13	5	37
2	L	19	33	14	20	
2	G	21	15	16	15	

2	S	5	10	5	0	
2	O	5	1	8	-3	
3	C	16	10	12	20	82
3	L	27	39	15	56	
3	G	23	10	22	3	
3	S	6	13	3	4	
3	O	2	2	3	-1	
4	C	15	2	7	40	17
4	L	12	21	19	-28	
4	G	6	13	6	0	
4	S	6	16	4	4	
4	O	2	0	1	1	
5	C	5	1	3	10	23
5	L	20	21	12	32	
5	G	3	2	5	-6	
5	S	5	22	11	-12	
5	O	0	0	1	-1	
6	C	6	1	3	15	47
6	L	13	21	9	16	
6	G	9	3	5	12	
6	S	9	10	6	6	
6	O	1	4	3	-2	
7	C	32	8	20	60	105
7	L	21	41	18	12	
7	G	39	2	29	30	
7	S	3	16	2	2	
7	O	1	0	0	1	

Tabla 10: Impacto de mejora por texto

Se puede apreciar que los textos con más cambios, que son el texto 7 y el texto 3, en este orden, también son aquellos que lideran el ránking de impacto. No se puede decir lo mismo de la parte baja de la tabla, pues el texto con menor índice de impacto es el texto 4, un texto con un valor medio en lo que a la frecuencia de la ocurrencia de cambios se refiere. Sin embargo, este texto se ve fuertemente perjudicado por el índice negativo en la categoría L, lo que atribuimos al gran número de elementos relacionados con aspectos culturales del lugar de origen. Asimismo, encontramos un valor llamativamente bajo en el texto 5 donde es particularmente notable el comportamiento de la categoría S. Como este texto comprende muchas expresiones utilizadas para dirigirse a un usuario, se observa una serie de

expresiones inaceptables en las traducciones relacionadas con convenciones lingüísticas, como, por ejemplo, cuestiones de elección de la forma de cortesía o formas más familiares.

3.5 Limitaciones del estudio

Aunque los resultados del presente estudio son, sin lugar a dudas, muy interesantes y abarcan un aspecto poco estudiado, el presente trabajo también tiene algunas limitaciones que deben mencionarse. La primera consiste en que solo se ha estudiado una combinación lingüística y una dirección. Esto se debe, entre otras cosas, a que entendemos el presente estudio como una primera aproximación al tema. Sin embargo, para poder contrastar los resultados obtenidos, sería deseable poder llevar a cabo estudios similares más amplios que engloben más combinaciones lingüísticas y material textual más extenso.

La otra limitación importante del estudio radica en el hecho de que toda la extracción, asignación y valoración se realizó por una sola persona aunque se trataba de una traductora y docente muy experimentada. El hecho de que fuera siempre la misma persona, por un lado, es garantía de alcanzar un alto grado de coherencia, por otro, sin embargo, conlleva una mayor subjetividad a la hora de valorar los cambios.

4. Conclusiones

El objetivo principal de este estudio consistía en analizar la evolución de dos sistemas abiertos de TA neuronal para comprobar si las traducciones producidas por ellos mejoran con el tiempo. Los datos generales que hemos extraído del análisis parecen confirmar este postulado, dado que el número de cambios a mejor supera considerablemente los cambios a peor (35,68 % vs. 27,8 %).

No obstante, el estudio proporciona también otros datos destacables acerca de la evolución de estos sistemas. Es, por ejemplo, muy llamativo que la evolución hacia una mejora no es lineal, sino que se puede constatar un empeoramiento significativo de la calidad entre los primeros dos puntos de medición. Esto, junto al alto número de ítems extraídos, indica una gran inestabilidad y fluctuación de estos sistemas.

Asimismo, la comparación entre los dos sistemas ha arrojado datos interesantes. El número total de ítems extraídos es considerablemente menor

en DeepL, lo que parece indicar una mayor estabilidad. Al mismo tiempo, no se han encontrado diferencias significativas entre los dos sistemas con respecto a la valoración de los cambios. Sin embargo, sí se aprecian diferencias significativas en relación a las categorías que parecen indicar una mayor estabilidad a nivel morfosintáctico por parte de DeepL.

La evolución de las traducciones realizadas por la TA neuronal a lo largo del tiempo es un tema poco estudiado que merecería más atención y podría enfocarse tanto desde la perspectiva del usuario como también desde la perspectiva del desarrollador de sistemas y es, por lo tanto, un tema que podría impulsar colaboraciones más estrechas entre traductores/traductólogos y desarrolladores.

El presente estudio no es más que una primera incursión en un terreno en el que queda mucho por explorar y que ofrece un marco excelente para estudios interdisciplinarios o multilingües. Esperamos que tales estudios vean la luz en un futuro cercano.

Referencias bibliográficas

Baureithel, U. (2019). "Künstliche Intelligenz verdrängt den Übersetzer – und nicht nur ihn" < https://www.tagesspiegel.de/kultur/sprachberufe-in-gefahr-kuenstliche-intelligenz-verdraengt-den-uebersetzer-und-nicht-nur-ihn/25184534.html>. Fecha de consulta de la página: 26.03.20.

Casacuberta, Nolla, F.; Peris Abril, A. (2017). "Traducción automática neuronal", *Tradumàtica. Tecnologies de la Traducció*, 15, 66-74, DOI: 10.5565/rev/tradumatica.203.

Caswell, I.; Liang, B. (2020). "Recent Advances in Google Translate", entrada de Google AI Blog del 20 de junio de 2020 <https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html>. Fecha de consulta de la página: 21.08.20.

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*, 108,109–120. DOI: 10.1515/pralin-2017-0013.

Chung, H. (2018). "Wie übersetzt NMT Eigennamen und Zahlen?", *Lebende Sprachen*, 63/1, 142-167. DOI: https://doi.org/10.1515/les-2018-0007

Elezaj, R. (2019). "Will Technology Leave Human Translators Jobless?" < https://www.wethetalent.co/data-robots-ia/will-technology-leave-human-translators-jobless/>. Fecha de consulta de la página: 26.03.20.

Hassan, H. et. al. (2018). "Achieving Human Parity on Automatic Chinese to English News Translation"<https://arxiv.org/pdf/1803.05567.pdf>. Fecha de consulta de la página: 16.08.20.

Koehn, Philipp (2020). *Neural Machine Translation*. Cambridge: University Press. DOI: https://doi.org/10.1017/9781108608480

Läubli, S; Castilho, S.; Neubig, G.; Sennrich, R.; Shen, Q.; Toral, A. (2020). "A Set of Recommendations for Assessing Human–Machine Parity in Language Translation" < https://jair.org/index.php/jair/article/view/11371/26573>. Fecha de consulta de la página: 16.08.20. DOI: 10.1613/jair.1.11371

Le, Q,; Schuster, M. (2016). "A Neural Network for Machine Translation, at Production Scale" <https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html>. Fecha de consulta de la página: 29.03.20.

Lommel, A.; Burchardt, A.; Görög, A; Uszkoreit, H; Melby, A.K. MQM: "Multidimensional Quality Metrics (MQM) Issue Types" <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html>. Fecha de consulta de la página: 29.03.20.

López-Pereira, A. (2019). "Traducció automàtica neuronal i traducció automàtica estadística: percepció i productivitat", *Tradumàtica. Tecnologies de la Traducció,* 17, 1-19. DOI: https://doi.org/10.5565/rev/tradumatica.235.

Marr, B. (2018). "Will Machine Learning AI Make Human Translators An Endangered Species? < https://www.forbes.com/sites/bernardmarr/2018/08/24/will-machine-learning-ai-make-human-translators-an-endangered-species/#79df46933902>. Fecha de consulta de la página: 26.03.20.

Mair, N.; Schmidhofer, A. (2019). "How does NMT deal with gender?" Translation Technology in Education – Facilitator or Risk?, July 5, 2019, Nottingham (UK) < https://www.nottingham.ac.uk/conference/fac-arts/clas/translation-technology-in-education%E2%80%93facilitator-or-risk/documents/translation-technology-programme-2019.pdf>. Fecha de consulta de la página: 29.03.20.

Plass-Fleßenkämper, B. (2017). "DeepL aus Deutschland könnte Google Translate den Rang ablaufen" <https://www.wired.de/collection/tech/deepl-google-translate-linguee>. Fecha de consulta de la página: 29.03.20.

Sánchez Ramos, M.; Rico Pérez, C. (2020). *Traducción Automática*. Granada: Comares.

Shterionov, D.; Nagle, P.; Casanellas. L.; Superbo, R.; O'Dowd, T. (2018). "Human vs Automatic Quality Evaluation of NMT and PBSMT", *Machine Translation*, 32, 217-235. DOI: 10.1007/s10590-018-9220-z.

Vaswani, A. et al. (2017). "Attention Is All You Need" <https://arxiv.org/pdf/1706.03762.pdf>. Fecha de consulta de la página: 16.08.20.

Wu, Y; Schuster, M; Chen, Z; Le, Q.V.; Norou, M. (2016). "Google's Neural Machine Translation System: Bridging the Gapbetween Human and Machine Translation" < https://arxiv.org/pdf/1609.08144.pdf>. Fecha de consulta de la página: 16.08.20.