1611 REVISTA DE HISTORIA DE LA TRADUCCIÓN
A JOURNAL OF TRANSLATION HISTORY
REVISTA D'HISTÒRIA DE LA TRADUCCIÓ

NÚMEROS | ISSUES | NÚMEROS | ARTÍCULOS | ARTICLES | ARTICLES | ESCENARIOS | SCENES | ESCENARIS

NÚMERO 19
ISSUE 19

LA TRADUCCIÓN AUTOMÁTICA EN EL SIGLO XXI

Núria BEL

Departamento de Traducción y Ciencias del Lenguaje
Universitat Pompeu Fabra

2025

Recibido: 15 de septiembre de 2025
Aceptado: 1 de noviembre de 2025

1. Introducción

Los resultados de los modelos grandes de lenguaje en el marco de la llamada inteligencia artificial generativa han admirado al gran público a pesar de que la historia de estos ingenios empezó alrededor de 1950, cuando Warren Weaver creyó saber cómo hacer una máquina de traducción automática. La investigación y la disponibilidad de ordenadores con cada vez más capacidad de computación fueron clave para que a principios del siglo XXI la traducción automática fuera ya un producto de consumo masivo. Por ejemplo, en 2006, Google lanza su traductor automático y de uso gratuito *Translate*, un sistema todavía basado en reglas que era la tecnología en la que estaban basados los sistemas profesionales que ya se venían utilizando en sectores de alta demanda de traducción como la ONU, la entonces Comunidad Económica Europea, o la Organización Panamericana de la Salud. A partir de aquí se suceden rápidamente las novedades que llevan hasta la aparición de ChatGPT. Google en 2007 lanza un nuevo traductor que utiliza tecnología estadística, y en 2016 otro con tecnología neuronal. En ese año, Google publica en su blog oficial que traduce 100.000.000.000 palabras al día (Turovsky, 2016). Pero ya no es el único gran proveedor de traducción, quizá solamente el más utilizado por el gran público. Otras compañías como DeepL, Amazon, Microsoft, PROMT, SDL, Baidu, y Alibaba tienen también productos de traducción automática, aunque algunos dirigidos específicamente a la traducción profesional. Durante los últimos años, la traducción entre algunos pares de idiomas ha mejorado sensiblemente, como se puede se puede ver en el Ejemplo 1 de traducción de inglés a español. A finales de 2022, OpenAI dio a conocer al gran público ChatGPT que, entre otras tareas, que también traducía, con tanta solvencia como ya lo hacía la traducción neuronal.

Ejemplo 1. Mejora de la traducción

Frase original	*They may not be exploited, reproduced, distributed, modified, publicly communicated, ceded or transformed.*

Google, 2016	Ellos no pueden ser explotados, reproducción, distribución, modificación, comunicación, cesión o transformación públicamente.
Google, 2021	No pueden ser explotados, reproducidos, distribuidos, modificados, comunicados públicamente, cedidos o transformados.
ChatGPT, 2025	No podrán ser explotados, reproducidos, distribuidos, modificados, comunicados públicamente, cedidos ni transformados.

La naturaleza y cantidad de los datos utilizados para desarrollar traducción automática es una, quizá la más importante, diferencia entre las diferentes tecnologías y sistemas. Los sistemas de traducción basados en reglas, la tecnología del siglo XXI, eran enormes compilaciones de condiciones sobre la equivalencia de dos expresiones, cada una en un idioma. Las condiciones se expresaban como conocimiento lingüístico, y había programas específicos que inducían ese conocimiento formulado, a su vez, con reglas. Por ejemplo, el conocimiento utilizado sería del tipo «el verbo inglés *to work* se traduce por *trabajar* si el sujeto es animado, y por *funcionar* si el sujeto es inanimado». Un analizador sintáctico automático tenía que identificar qué palabras eran el sujeto de la oración, conocimiento que también se formulaba en términos de declaraciones de condiciones, y buscar en una especie de diccionario con información semántica sobre cada palabra, si el nombre que era el núcleo del sujeto correspondía a una entidad animada o inanimada.

Además de requerir de mucha posedición, porque las traducciones de los sistemas basados en reglas eran muy literales, el gran problema de esta tecnología era que cada par de traducción implicaba la redacción de miles de reglas, con condiciones sobre el contexto, y de diccionarios con un gran número de características lingüísticas y extralingüísticas de las palabras. Para crear y mantener cada nuevo par de lenguas eran necesarios muchos lingüistas y traductores, y, sobre todo, tiempo. Así, a finales del siglo XX y principios del XXI, la llamada traducción automática estadística fue considerada una mejor opción, no porque se consiguiera una mejora de la calidad de la traducción, sino especialmente porque reducía el coste del desarrollo de nuevos pares. Los sistemas estadísticos no requerían lingüistas o traductores, ni siquiera necesitaban diccionarios bilingües. Solo era necesario disponer de textos originales y sus traducciones. Lo mismo que necesitan los sistemas actuales: los neuronales y los modelos grandes del lenguaje, base de la inteligencia artificial generativa.

En este artículo nos proponemos explicar la tecnología que puede dar lugar a un sistema de traducción automática sin expertos en traducción, ni en lenguas. La explicación quiere facilitar una comprensión intuitiva, concentrándose en los detalles que afectan a la tarea de traducir, prescindiendo de los que se refiere a la ingeniería informática necesaria para hacer los millones de millones de cálculos necesarios. Con esta explicación aspiramos a que los profesionales de la traducción puedan evaluar las capacidades de estas tecnologías y así entender sus limitaciones, y, más importante, que conozcan su absoluta dependencia de las traducciones ya existentes, y cómo la cantidad disponible de estas determinan inexorablemente la calidad de las traducciones producidas.

Sin embargo, nuestra explicación de los desarrollos tecnológicos que son la historia más reciente de la traducción automática se ve limitada por el hecho de que estudiamos productos comerciales y no toda la información técnica es pública. Nuestras explicaciones están basadas en los artículos académicos y otros documentos publicados por los desarrolladores de estos sistemas antes del producto comercial. Además, hemos seleccionado solamente las publicaciones que marcaron la evolución de esos sistemas y que son la historia de la tecnología.

En lo que sigue, y después de esta introducción, en la sección 2 presentamos las bases de la tecnología estadística de traducción automática, cuyas innovadoras propuestas sirvieron de base para la tecnología neuronal que presentamos en la sección 3. En la sección 4, explicamos cómo se consigue que los asistentes basados en modelos grandes del lenguaje puedan resolver tareas de traducción, entre otras como el resumen automático, la respuesta a preguntas, etcétera. Acabamos este artículo con unas observaciones finales que forman la sección 5.

2. Traducción automática estadística (TAE)

Brown *et al*. (1990) fue la presentación académica de una innovadora forma de crear las correspondencias entre palabras y expresiones en dos lenguas diferentes a partir de textos y traducciones ya existentes. La solución propuesta era posible gracias a la creciente capacidad de computación de los ordenadores de la época. Con esas nuevas capacidades, los programas diseñados para convertir un input en un output podían basarse en realizar millones de veces una operación. Se podían encadenar varias operaciones sencillas en un algoritmo y hacer que la máquina las ejecutara millones de veces, sin apenas coste adicional y rápidamente. De esta forma, soluciones que parecían demasiado costosas o incluso imposibles si debían realizarlas trabajadores humanos, se convertían en viables.

Para traducir, la idea básica era que, en cada par de frases, una siendo la traducción de la otra, cualquier palabra de la frase meta podía ser la traducción de cualquier palabra de la frase original, pero que, si se disponía de un gran número de pares de frases, el número de veces que resultaría la traducción correcta sería mayor que los emparejamientos incorrectos. Es decir, a partir de un corpus paralelo (nombre que se daba a la colección de pares de frases y sus traducciones) como el Ejemplo 2 se emparejaban todas con todas, para ver qué emparejamiento resultaba más frecuente y que era propuesto como correspondencia de traducción. Como vemos en el sencillo Ejemplo 2, la hipótesis de que la correspondencia correcta resulta del emparejamiento más frecuente se suele cumplir.

Ejemplo 2. Emparejamientos y frecuencia

Corpus paralelo

a. The fast elevator works = El ascensor rápido funciona.
b. This fast elevator works = Este ascensor rápido funciona.
c. The elevator is closed = El ascensor está cerrado.

Emparejamientos

Palabra original	Palabra traducción	Número de veces
the	el	2
the	ascensor	1
the	rápido	2
the	funciona	2
fast	el	2
fast	ascensor	2
fast	rápido	2

the	funciona	2
fast	funciona	2
elevator	el	2
elevator	**ascensor**	**3**
elevator	funciona	2
this	este	1
the	funciona	2
this	este	1
this	ascensor	1
this	rápido	1
this	funciona	1
is	el	1
is	ascensor	1
is	está	1
is	cerrado	1
closed	el	1
closed	ascensor	1
closed	está	1
closed	cerrado	1

El par *the-ascensor* sería menos frecuente que *elevator-ascensor*, que resultaría seguro de cada par de frases en las que aparezcan estas palabras. *the-ascensor* aparecería en muchos pares, pero no tantos, porque habría frases con expresiones como *this elevator*, *an elevator*, etcétera. Además, podía calcularse la traducción más frecuente entre varias posibles: *elevator-ascensor* pero también pares como *elevator-elevador*, y *elevator-montacargas*. A partir de los datos del corpus, se podía calcular que en un 80% de casos la correspondencia encontrada era *elevator-ascensor*, en un 15% *elevator-montacargas* y en un 5% *elevator-elevador*, por ejemplo. Pero, para obtener esas correspondencias, la cantidad y calidad de datos era crucial. Los experimentos de Brown *et al*. (1990) mostraron que sólo era posible por la disponibilidad de 1.778.620 pares de frases inglés-francés: el corpus de traducciones oficiales de las sesiones del Parlamento de Canadá (Hansards 1974-1978).

El cálculo de la probabilidad de las correspondencias entre palabras se conoce desde entonces como el **modelo de traducción** que una vez entrenado, es decir cuando a partir de los datos se ha calculado la probabilidad de las correspondencias entre todas las palabras de los textos, se puede utilizar para predecir la traducción de una frase nueva, generándola a partir de la selección de la correspondencia más probable de las palabras de la frase original.

Las correspondencias entre las palabras podían estar bien, pero también tenían que encontrar la forma de ordenarlas según la lengua meta. De la tarea de la fluidez en

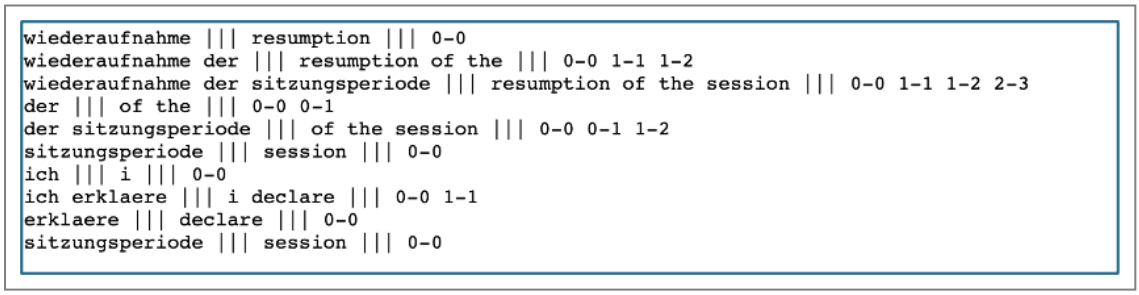
la lengua meta se encargaba otro componente estadístico, un **modelo de lenguaje**, con el que se estimaba la probabilidad de las secuencias de palabras en un idioma determinado. Tras la combinación en todos los órdenes posibles de las palabras resultado de aplicar el modelo de traducción, el modelo de lenguaje calculaba qué frase resultaba la más probable en la lengua meta, como en el Ejemplo 3. Este modelo del lenguaje es, efectivamente, un precursor de los modelos grandes de lenguaje (*Large Language Models*, LLM) que son la base de la inteligencia artificial generativa actual.

Ejemplo 3. Diferentes salidas del modelo de traducción para *The fast elevator works*

a. El rápido ascensor funciona.
b. El ascensor rápido funciona.
c. El ascensor funciona rápido.

El entrenamiento del modelo de lenguaje es el cálculo de probabilidades de secuencias de palabras a partir de muchos textos de la lengua modelada. Esas probabilidades se usarán en producción para calcular la frase meta más probable de entre todas las posibles traducciones de una frase original nueva proporcionadas por el módulo de traducción. Como ya hemos dicho, esa forma nueva de abordar la traducción, sin intervención de traductores ni lingüistas, se basaba crucialmente en dos recursos cada vez más disponibles en el siglo XXI: gran cantidad de textos traducidos y digitalizados y ordenadores que pudieran hacer rápidamente todos esos cálculos, sencillos, pero millones de millones.

La traducción automática estadística evolucionó entonces a buscar la traducción en grupos de más de una palabra (Traducción automática estadística basada en frases) para elegir una u otra traducción. Koehn *et al*. (2003) presentaron una solución, no muy sofisticada desde el punto de vista del traductor, aunque aumentaba la cantidad y complejidad práctica de los cálculos en el entrenamiento. Decidieron buscar correspondencias por grupos de 1, 2 y 3 palabras. Alargaron el número de elementos de las correspondencias para intentar listar y así almacenar la traducción según las palabras circundantes. Las tablas de correspondencias bilingües eran como la de la Figura 1, que está copiada de uno de los cursos de iniciación del sistema presentado en Koehn *et al*. (2007), MOSES.**(1)**



```
wiederaufnahme ||| resumption ||| 0-0
wiederaufnahme der ||| resumption of the ||| 0-0 1-1 1-2
wiederaufnahme der sitzungsperiode ||| resumption of the session ||| 0-0 1-1 1-2 2-3
der ||| of the ||| 0-0 0-1
der sitzungsperiode ||| of the session ||| 0-0 0-1 1-2
sitzungsperiode ||| session ||| 0-0
ich ||| i ||| 0-0
ich erklaere ||| i declare ||| 0-0 1-1
erklaere ||| declare ||| 0-0
sitzungsperiode ||| session ||| 0-0
```

Figura 1. Ejemplo de tabla de correspondencias extraídas automáticamente por el sistema de traducción automática estadística basada en frases MOSES (Koehn *et al*., 2007). Cada línea corresponde a un segmento de la lengua fuente, en este caso alemán, la frase correspondiente en la lengua meta, aquí inglés, y la correspondencia de las palabras expresada en la posición y que empieza por 0. Por ejemplo, en varios casos la palabra *der* se alinea con dos palabras del inglés: 1-1 y 1-2.

La traducción automática estadística primero y la traducción automática estadística basada en frases después fueron mejorando la calidad de las traducciones, pero sobre todo redujeron la cantidad de esfuerzo humano necesario para desarrollar un nuevo par de lenguas o un nuevo dominio, aunque ahora dependían de la cantidad de textos traducidos disponibles para cada nuevo par de lenguas (o para un ámbito de conocimiento o dominio específico). Si no se disponía de un gran corpus paralelo, original y traducción, no se podían extraer las correspondencias, así que por ejemplo Google Translate iba añadiendo pares de lenguas a su lista de traductores utilizando alguna lengua (normalmente el inglés) como interlingua. Por ejemplo, traducía del rumano al inglés y del inglés al castellano para poder traducir del rumano al castellano. O hubo incluso pares de lenguas que se resolvían con tres traductores: catalán-castellano, castellano-inglés, y del inglés a cualquier otra del catálogo. El

usuario no veía las traducciones intermedias, aunque a menudo notaba los errores que se creaban en una de las traducciones intermedias y que se arrastraban hasta la traducción final.

La tecnología estadística de Brown *et al*. (1990) y Koehn *et al*. (2003) fue clave para destacar el papel del modelo de lenguaje en la traducción y para definir el problema de la traducción como el problema de encontrar la secuencia de palabras más probable dadas las palabras de una frase en otro idioma. Esta definición será la clave de las innovaciones aportadas en la traducción automática neuronal y de que los modelos grandes de lenguaje puedan traducir y hacer otras tareas como resumir, corregir, contestar a preguntas, etcétera.

3. La traducción automática neuronal (TAN)

A finales del siglo xx ya se había considerado la posibilidad de utilizar la tecnología de aprendizaje automático llamada 'redes neuronales' para resolver la tarea de la traducción (por ejemplo, Forcada y Ñeco, 1997; Castaño y Casacuberta, 1997), pero fue el gran aumento en potencia de cálculo de los ordenadores de los siguientes quince años lo que lo hizo posible. La primera tecnología llamada 'secuencia a secuencia' (Cho *et al*., 2014; Sutskever *et al*. 2014) y la inmediatamente posterior llamada 'transformador' (Vaswani *et al*., 2017) mejoraron las traducciones pero a costa de hacer millones de millones de operaciones de cálculo.

Una red neuronal es un entramado de unidades de computación a las que llaman neuronas por analogía con las células del cerebro que inspiraron esta tecnología de aprendizaje automático (Mitchell, 1997). Las neuronas son unidades de computación porque cada una de ellas recibe unos cuantos valores de entrada y produce un único valor de salida. Cuando se organizan en una red, el resultado de un número de neuronas de una primera capa se convierte en el *input* o valores de entrada de las neuronas de la capa siguiente, como se muestra en la Figura 2. El resultado de la segunda capa, a su vez, serán los valores de entrada de la siguiente capa, y así se va procesando de forma cada vez más profunda. Las conexiones entre las neuronas (los parámetros) son ponderables: cada conexión tiene un peso que determina cómo de importante es esa computación para el resultado deseado. El entrenamiento sigue siendo calcular las probabilidades de palabras que generarían de nuevo el corpus usado como datos.

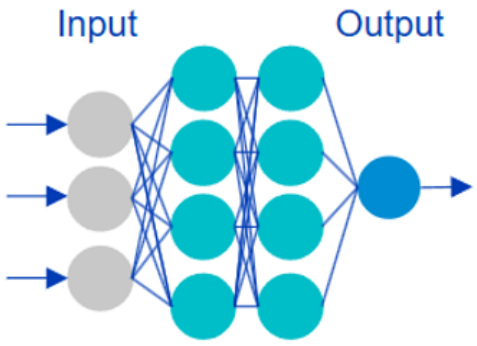


Figura 2. Representación esquemática de una red neuronal. A partir de varios inputs produce un único resultado: dadas unas palabras, qué palabra es la más probable como continuación. Elaboración propia

Por ejemplo, en un modelo de lenguaje el objetivo del entrenamiento es volver a producir (o casi) los datos del corpus de entrenamiento. Ha de calcular la probabilidad que tiene cada palabra del vocabulario después de las palabras anteriores. El programa va probando pesos y comparando la propuesta de palabra siguiente con la que hay en el corpus. Cuando no acierta, reasigna pesos y vuelve a intentarlo hasta que da con los pesos de todas las conexiones de la red que generan el texto que más se parece al texto de entrenamiento, o hasta que el coste o el tiempo de computación se considera aceptable. Por eso se puede decir que aprende,

porque va cambiando el peso de cada conexión, y hay millones, hasta que generen lo que hay en el corpus de entrenamiento. Entrenar un modelo de lenguaje suele tardar varios meses, incluso usando decenas de miles de potentes GPU.

Como en la traducción automática estadística, un modelo de lenguaje calcula la probabilidad de una palabra dadas las anteriores en la lengua meta, pero la novedad de la traducción neuronal es que la probabilidad de la siguiente palabra está condicionada además por las palabras de la frase original. Como se ve en la Figura 3, un módulo llamado codificador procesa las palabras de la frase original hasta obtener una representación profunda sobre la que ha de condicionar la generación de palabras hasta completar la frase que va a ser la traducción.

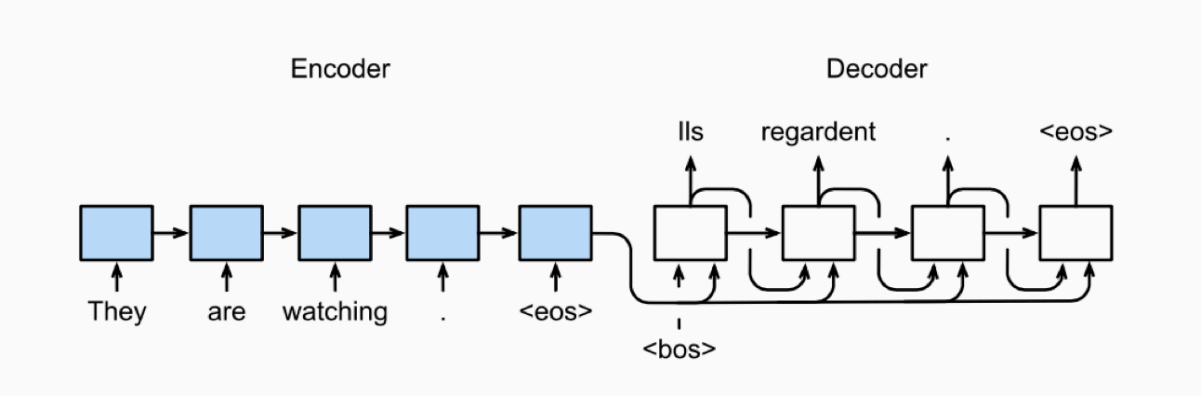


Figura 3. Representación tomada de Wikimedia Commons de un modelo secuencia a secuencia para traducir. A la máquina que construía esa representación de la frase del idioma origen se la llamó «codificador» *(encoder)* y entonces la máquina que iba generando la traducción a partir de la representación proporcionada se la llamó «decodificador» *(decoder)*.
Créditos: Zhang, Lipton, Zachary, Li y Smola. CC BY 4.0

Una de las claves de la tecnología neuronal fue que la representación de palabras y frases dejaban de ser cadenas de caracteres alfabéticos, símbolos. Con las redes neuronales las palabras se representan con un vector numérico, una lista ordenada de números sin ningún significado, pero diferente para cada palabra (Bengio *et al*., 2003). En el procesamiento que se produce durante el entrenamiento, la representación de cada palabra se va enriqueciendo cuando se le añaden vectores de algunas de las palabras con las que coaparece.

Para explicarlo de forma intuitiva podemos decir que al principio una palabra como 'works' se representa con un vector [1,2,3,4] y que 'elevator' tiene la representación [5,6,7,8]. El codificador se encarga de que, después de ver diferentes ejemplos, 'works' tenga la representación [1,2,3,4,5,6,7,8] que es la relacionada con la traducción 'funciona'. También después de ver otros ejemplos, se crea una representación de 'works' [1,2,3,4,9,10,11,12], porque ha añadido la información de palabras con las que se combina frecuentemente como 'boy', 'man', etc, que se representan como [9,10,11,12]. Se dice que esta representación vectorial codificada es semántica en el sentido de la hipótesis distribucional de Firth (1954) por la que se considera que palabras que aparecen en los mismos contextos tienden a tener significados relacionados. Como acabamos de ver, en el caso de las representaciones contextuales, palabras diferentes pero que se combinan frecuentemente con las mismas palabras se parecerán y se podría inferir que tienen una semántica parecida. 'The lift works' se traducirá por 'El ascensor funciona' aunque esta correspondencia no esté en el corpus, porque los contextos en los que aparecen 'lift' y 'elevator' son muy parecidos, y los vectores contextuales se parecerán. Para poner un ejemplo más, en las primeras versiones de la NMT, se daban con cierta frecuencia errores debidos a esta transferencia entre palabras de significado y contexto habitual similar pero que no se correspondían a lo que se decía en el original. Por ejemplo, predecir *Noruega*, cuando el original habla de Túnez (Arthur *et al*. 2016): 'I come from Tunisia' = 'Vengo de Noruega'.

Otro elemento clave que se introdujo con la traducción neuronal fue procesar y representar trozos de palabras, lo que llaman *subpalabras*. Uno de los problemas de trabajar con probabilidades es que, como siempre se refiere a la frecuencia con que

se ha observado una determinada palabra, si en el corpus de entrenamiento no estaba esa palabra, la probabilidad calculada es cero. Dividiendo en uno o varios segmentos las palabras con un algoritmo específico (Byte-Pair Encoding, BPE, Senrich *et al*., 2016) se podía asegurar que los trozos de palabras resultantes sí que iban a estar en el corpus. Los *tokenizers*, como se llaman a estos algoritmos que identifican trozos de palabras, no tienen nada que haga que la partición se parezca a raíz y sufijos, así que podría haber subpalabras como 'asc', 'en' y 'sor'. El cálculo de la probabilidad de la siguiente palabra, se convierte en la probabilidad de la siguiente subpalabra, lo que puede dar lugar a que genere secuencias de caracteres que no son palabras de la lengua de llegada, aunque podría haberlo sido, como en el ejemplo de la Figura 4.

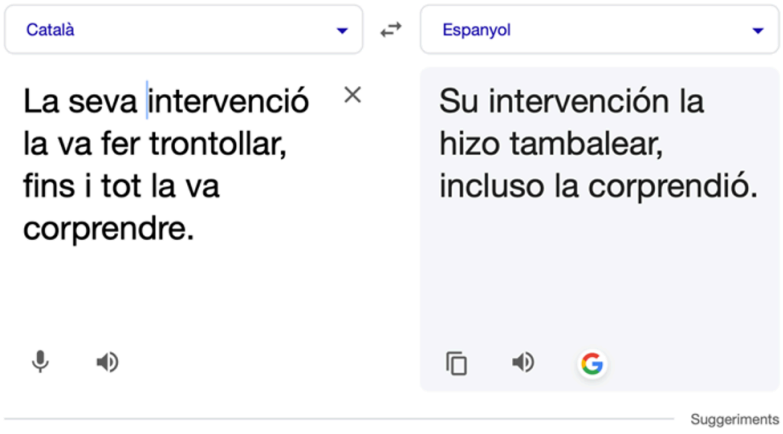


Figura 4. Captura de pantalla con ejemplo de traducción con Google Translate que genera una palabra bien flexionada 'corprendió' pero que no existe y que es una consecuencia de la segmentación en subpalabras

Por último, para mejorar la calidad, la TA neuronal emplea un método específico conocido como *beam search* (traducido normalmente por 'búsqueda por haz'). Este método mantiene activas varias hipótesis de traducción (cada hipótesis con la *n* palabra más probable) producidas por el decodificador hasta que se alcanza el final del segmento a traducir. En ese punto, el sistema evalúa todas las secuencias producidas para determinar la secuencia más probable en conjunto (Koehn *et al*., 2003; Freitag y Al-Onaizan, 2017).

Con todas estas innovaciones, la TA neuronal mejoró sensiblemente los resultados de la traducción automática estadística, pero necesitaba de muchísimos más datos. Koehn y Knowles (2017) compararon la calidad de la traducción de dos sistemas inglés-español, uno estadístico basado en frases y otro neuronal, entrenados ambos con diferentes cantidades de datos de entrenamiento: de 400.000 palabras a 385 millones de palabras. El sistema neuronal solamente empezaba a ir mejor que el estadístico a partir de 15 millones de palabras, aunque iba mejorando muy rápidamente conforme se iba aumentando esta cifra. Esos autores ofrecían ejemplos de la mejora de la calidad al ir aportando más datos. Reproducimos sus resultados en la Tabla 1. Esta dependencia de la cantidad de datos va a ser aun mayor en la siguiente tecnología, la inteligencia artificial generativa.

Número de palabras del corpus de entrenamiento	Frase original: *A Republican strategy to counter the re-election of Obama*
376.000	Un órgano de coordinación para el anuncio de libre determinación
753.320	Lista de una estrategia para luchar contra la elección de hojas de Ohio
1.500.000	Explosión realiza una estrategia divisiva de luchar contra las elecciones de autor

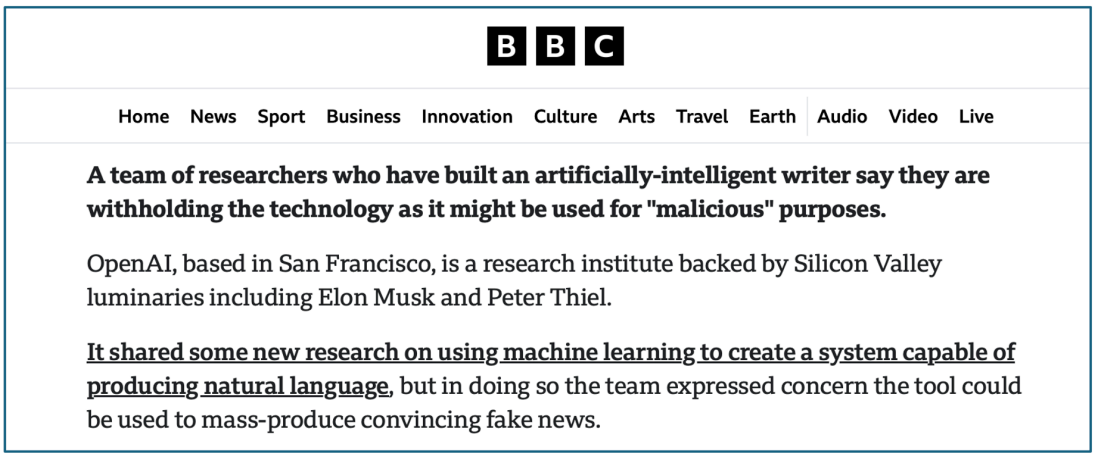
6.026.562	Estrategia siria para contrarrestar la reelección del Obama
12.053.125	Una estrategia republicana para contrarrestar la reelección de Obama

Tabla 1. Traducción de la primera frase de los datos de prueba usando un sistema" de traducción neuronal entrenado con diferentes cantidades de datos. Tomado de Koehn y Knowles (2017), quienes constataban que, con pocos datos de entrenamiento, la traducción neuronal producía resultados que, a pesar de ser frases posibles en la lengua de llegada, no tenían relación alguna con la frase original

4. Inteligencia artificial generativa para la traducción

Los modelos grandes de lenguaje (a partir de ahora usaré el acrónimo LLM, del inglés Large Language Models) con arquitectura GPT (del inglés, Generalised Pre-Trained Transformers) son decodificadores, como los que acabamos de presentar. Generan texto al calcular la palabra más probable dadas otras palabras anteriores, pero sin codificador, usando los cálculos realizados en el entrenamiento con datos. Esas palabras anteriores suelen ser las de la instrucción o comando que da un usuario (el *prompt*, en inglés) y las que, de forma iterativa, va generando el mismo sistema.

El primer modelo GPT que fue noticia en 2019 tenía unos resultados bastante modestos, pero OpenAI, la compañía que lo desarrollaba, sorprendió al publicar que la nueva tecnología era 'peligrosa' para el mundo y dejó de hacer público el código (Figura 5). Con el modelo siguiente, GPT3 (Brown *et al*., 2020), los resultados mejoraron. GPT3 era mucho mayor que los modelos anteriores en el número de parámetros (las conexiones entre neuronas) y en la cantidad de datos utilizados para el entrenamiento. Si GPT2 tenía 1.500 millones de parámetros, GPT3 tenía 175.000 millones. Para entrenar GPT2 habían utilizado textos con 40.000 millones de palabras provenientes de Internet, colecciones de libros, páginas de la Wikipedia; para GPT3 usaron 300.000 millones de palabras que, según la empresa, provenían de textos de las mismas fuentes y en un 93 % en inglés.


BBC

Home News Sport Business Innovation Culture Arts Travel Earth Audio Video Live

A team of researchers who have built an artificially-intelligent writer say they are withholding the technology as it might be used for "malicious" purposes.

OpenAI, based in San Francisco, is a research institute backed by Silicon Valley luminaries including Elon Musk and Peter Thiel.

It shared some new research on using machine learning to create a system capable of producing natural language, but in doing so the team expressed concern the tool could be used to mass-produce convincing fake news.

Figura 5. Captura de noticia publicada por la BBC comentando el comunicado de OpenAI sobre la peligrosidad de GPT3 en febrero de 2019. Fuente: **https://www.bbc.com/news/technology-47249163**

En Ouyang *et al*. (2022) se publicaron datos sobre el modelo InstructGPT, antecedente inmediato de ChatGPT, que tenía dos importantes novedades. Una era que los resultados del modelo de lenguaje, el texto generado palabra a palabra, podía ser controlado por un método que había sido entrenado para reconocer texto inapropiado. Construyeron un corpus de entrenamiento específico en el que diferentes personas habían puesto nota a los textos que generaba el modelo de lenguaje GPT2 que, quizá como consecuencia de las fuentes usadas, contenía con frecuencia muestras de racismo y machismo, así como un léxico a menudo procaz e inapropiado para convertirse en un producto comercial. Con esos nuevos datos,

añadieron un entrenamiento adicional para que el modelo aprendiera a clasificar los textos de acuerdo con las notas que habían puesto los anotadores humanos, de forma que se podía identificar mensajes no apropiados y descartarlos como respuesta. La otra novedad era que se dieron cuenta de que los textos con los que entrenar podían ser una secuencia de nombre o descripción de la tarea, texto de entrada y el resultado deseado. En el caso de la tarea de traducción, los datos tendrían el aspecto de la captura de pantalla de la Figura 6.

Figura 6. Ejemplo de datos de entrenamiento de un gran modelo de lenguaje para la tarea de traducción de inglés al búlgaro. Captura de pantalla del conjunto de datos Europa Education and Culture Translation Memory (EAC-TM) publicados en la plataforma HuggingFace.co.
Fuente: **https://huggingface.co/datasets/community-datasets/europa_eac_tm**

Con este tipo de datos se 'instruía' el sistema, en el sentido de que como se calculaba la probabilidad de las secuencias palabras, el modelo podía generar un nuevo texto que tuviera la secuencia más probable que era la solución a la tarea. Así se simulaba haber resuelto la tarea propuesta por el usuario.

Ahora, a partir del input del usuario, el *prompt*, el modelo genera la respuesta palabra a palabra basándose en la probabilidad calculada a partir de los datos de pre-entrenamiento, los del modelo de lenguaje que hemos visto, y refinarlos con los de instrucción. Si los textos de instrucción han sido pares de frases, una la traducción de la otra, el modelo acaba calculando la probabilidad de que las palabras en una lengua aparezcan detrás de la secuencia de palabras en otra lengua. La identidad de los idiomas se puede convertir en un condicionante más de la probabilidad de la siguiente palabra. Después de la frase en el Ejemplo 4, la palabra más probable será 'coffee'. Su probabilidad será más alta que la de la palabra 'tea' que en cambio será más probable cuando en la frase original aparezca 'té', y mucho más probable que la palabra 'Kaffee' del alemán, que solamente sería más probable si la frase del usuario tuviera la palabra 'alemán' y quizá otras como 'Ich mag'.

Ejemplo 4. Traduce al inglés: Me gusta el café. *I like...*

Es decir, para que un *prompt* como «Traduce esta frase al inglés: Me gusta el café», se pueda resolver con la información de qué palabra es la más probable detrás de estas otras, se han usado millones de frases en las que aparecen palabras como 'traduce', 'inglés', y frases originales y sus traducciones.

Se han recopilado ya millones de textos y su traducción a partir de libros, Wikipedia, otras páginas web multilingües y memorias de traducción existentes como la que hemos visto en la Figura 6. La traducción se hace de la misma forma que para todas las otras tareas a las que estos asistentes dan respuestas. Por ejemplo, el conjunto de datos Aya (Singh *et al.*, 2024) son 503 millones de instrucciones, en 114 lenguas diferentes, nombrando 12 tareas entre las que se incluyen respuesta a preguntas,

resumen, traducción, paráfrasis, análisis de opinión, e inferencia en lenguaje natural. Flan 2022 (Longpre *et al*., 2023) consiste en 15 millones de ejemplos de instrucciones para 1.836 tareas diferentes. También hay datos específicos para tareas como la corrección gramatical (Raheja *et al*., 2023) como los que vemos en la Figura 7.

Figura 7. Captura de pantalla del conjunto de datos utilizado por Grammarly para entrenar el LLM CoEdit (Raheja *et al*. 2023), publicado en la plataforma HuggingFace.co.
Fuente: **https://huggingface.co/datasets/grammarly/coedit**

La gran aportación de la IA generativa es toda la ingeniería necesaria para procesar estos textos con millones de millones de palabras, signos de puntuación, números, etcétera, y calcular rápidamente una probabilidad que permite generar secuencias de palabras condicionadas las unas a las otras. (Para una explicación más detallada, véase Bel, 2025.)

Otra de las grandes contribuciones de estos LLM es la cantidad de palabras que pueden tener en cuenta para generar la siguiente palabra. Como los LLM generan iterativamente palabras probables que siguen a las palabras anteriores, su contexto, cuantas más palabras haya en el *prompt*, la respuesta estará más relacionada y será, en principio, mejor. Por ejemplo, OpenAI anunciaba que GPT 4.1 nano **(2)** tenía la capacidad tener en cuenta un millón palabras, puntuación, símbolos, etc. de contexto, Gemini también anunció **(3)** que su sistema podía procesar como contexto 1.500 páginas de texto. Así, estos sistemas añaden por defecto los comandos y respuestas que se han ido generando en el curso del diálogo como contexto de los nuevos *prompts* para condicionar más la probabilidad y dar con las palabras que son realmente la respuesta. De hecho, algunos profesionales suelen recomendar, por ejemplo, incluir en el *prompt* ejemplos de traducciones que fuercen a que la respuesta contenga palabras del mismo dominio, con el mismo tono, etcétera. Se habla de *prompt engineering* para describir la heurística que conlleva encontrar instrucciones para conseguir la respuesta óptima a las tareas concretas propuestas por el usuario.

El estudio de Gao, Wang y Hou (2024) comprobaba la existencia de diferencias en la calidad de la traducción propuesta por ChatGPT al incluir en el *prompt* (i) información sobre la tarea, por ejemplo, mencionando los idiomas y la dirección de traducción (‘de inglés a español’) y (ii) información sobre el dominio para mejorar la terminología. Para evaluar, dichos autores utilizaron, entre otras medidas, la tasa de error de traducción (TER, del inglés *Translation Error Rate,* Snover *et al*., 2006) que mide los cambios que se han de realizar en la frase resultado del sistema para que se parezca a una traducción de referencia. Compararon los resultados de Google Translate, DeepL y ChatGPT obtenidos por medio de los *prompts* específicos que acabamos de mencionar con frases y sus traducciones extraídas de Wikipedia. Todas las pruebas se hicieron en la dirección inglés a otra lengua, y las pruebas se limitaron a 50 frases por lengua. Los resultados se muestran en la Figura 8, en la que se observa que los diferentes sistemas son bastante similares y que lo que tiene mayor impacto en la calidad de la traducción son los pares de lenguas. La importancia de la lengua origen y la lengua meta es en realidad la importancia de la cantidad de datos (Moslem *et al*., 2023; Robinson *et al*., 2023). Por ejemplo, Robinson *et al*. (2023), después de probar la traducción del inglés a otros 204

idiomas, demostraron que había una correlación positiva estadísticamente significativa entre la calidad de los resultados y el número de páginas publicadas en la Wikipedia en esas lenguas, independientemente de lo que se incluía en el *prompt*, aunque en Moslem *et al*. (2023) se muestra una ligera mejora en la terminología si se incluyen en el *prompt* algunos ejemplos de traducciones con esa terminología, pero, de nuevo, solo para los pares de lenguas para los que se ha dispuesto de más datos.

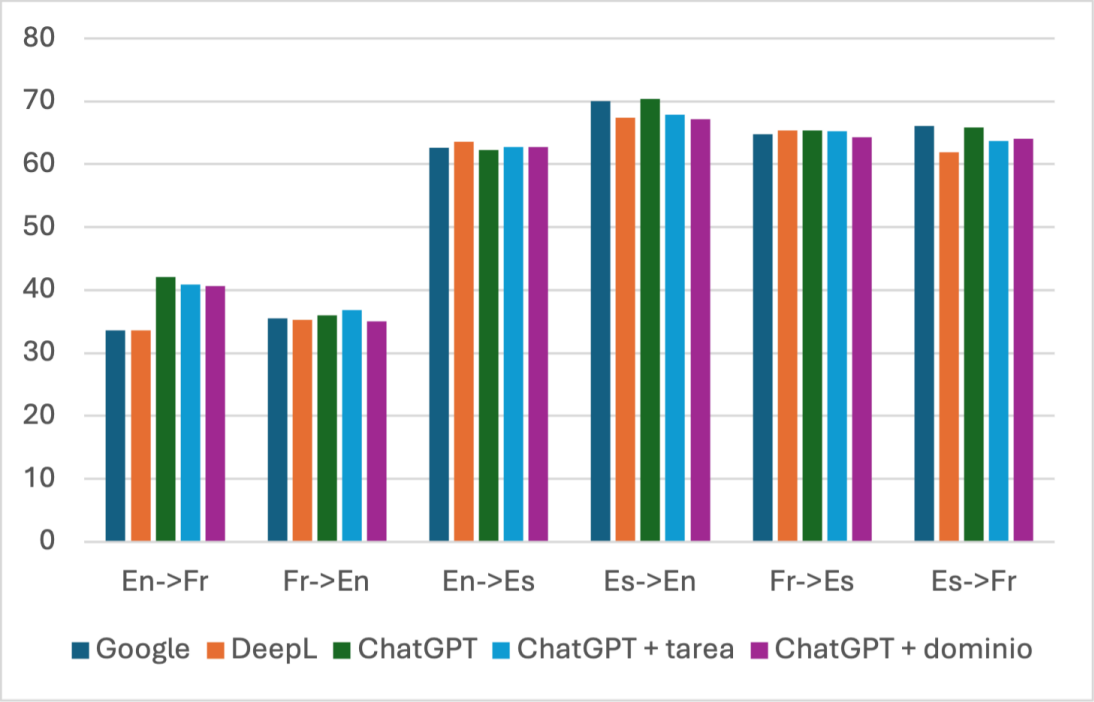


Figura 8. Resultados en términos de Tasa de error de traducción (TER, menor puntuación es mayor calidad) de diferentes sistemas de traducción neuronal (Google y DeepL) comparado con los resultados de ChatGPT obtenidos con diferentes *prompts*: sin instrucciones específicas con especificación del par y dirección de traducción, y con información adicional sobre el dominio. Según **https://meta.wikimedia.org/wiki/List_of_Wikipedias**, hay 64 millones de páginas en inglés, 13,6 millones en francés y 8,5 millones en español. Elaboración propia a partir de los datos de Gao, Wang y Hou (2023)

5. Observaciones finales

De esta breve explicación de la traducción automática del siglo XXI se puede entrever que la mejora de la calidad que se ha ido produciendo se debe sin ninguna duda a la utilización de cada vez más datos: de traducciones ya existentes. Hemos visto que se ha demostrado una correlación directa entre cantidad de datos en una lengua y la calidad de la traducción obtenida. Hemos intentado explicar de forma intuitiva que, dada la tecnología utilizada, no podría ser de otra manera ya que la información básica que manejan las diferentes tecnologías es la probabilidad de las secuencias de palabras y eso explica que la mejora de la calidad haya ido siempre de la mano de usar más datos, no de innovaciones que muestren un aumento de las capacidades de abstraer y generalizar.

Si tenemos en cuenta que, según Wikipedia, GPT4 ha sido entrenado con 13 millones de millones de palabras, código, etcétera, no es difícil deducir que ya han usado todos los textos disponibles y que mejorar la traducción entre algunos pares de lenguas ya no es posible. Algunos autores proponen la creación y uso de datos sintéticos para compensar la falta de datos, pero se ha demostrado que usar texto generado por modelos del lenguaje para entrenar afecta también la calidad de los resultados (Shumailov *et al*. 2024).

En conclusión, esta total dependencia de la cantidad de datos implica que los pares de lenguas con pocos textos traducidos disponibles no podrán obtener traducciones de calidad con la tecnología de lo que llevamos de siglo XXI.

NOTAS

(1) MOSES, statistical machine translation system, disponible en **https://www2.statmt.org/moses** [consultado el 29 de agosto de 2025].

(2) Open IA, «Introducing GPT-4.1 in the API», disponible en **https://openai.com/index/gpt-4-1/** [consultado el 29 de agosto de 2025].

(3) Gemini, «Profundiza en archivos grandes y repositorios de código», disponible en **https://gemini.google/overview/long-context/** [consultado el 29 de agosto de 2025].

BIBLIOGRAFÍA CITADA

BEL, Núria, «Un vocabulario básico de los grandes modelos del lenguaje: una explicación y 25 términos», *Terminàlia*, 1:31 (2025), 27–39, disponible en **https: //revistes.iec.cat/index.php/Terminalia/article/view/154678** [consultado el 29 de agosto de 2025].

BENJIO, Yoshua, *et al*., «A neural probabilistic language model», *Journal of Machine Learning Research*, 3 (2003), 1137–1155, disponible en **https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf** [consultado el 29 de agosto de 2025].

BROWN, P., *et al*., «A Statistical Approach to Language Translation» en *COLING '88: Proceedings of the 12th conference on Computational linguistics*, Stroudsburg (Pensilvania), Association for Computational Linguistics, 1988, pp. 71-76, disponible en **https://aclanthology.org/C88-1016.pdf** [consultado el 29 de agosto de 2025].

CASTAÑO, Asunción, *et al*.,«Machine Translation using Neural Networks and Finite-State Models», en *Proceedings of the 7th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, St John's College, Santa Fe, 1997.

CHO, Kyunghyun, *et al*., «On the Properties of Neural Machine Translation: Encoder–Decoder Approaches», en *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Association for Computational Linguistics, 2014, pp. 103-111.

FORCADA, Mikel L., y Ramón P. ÑECO, «Recursive Hetero-Associative Memories for Translation», en José Mira, Roberto Moreno-Díaz y Joan Cabestany (eds.), *Biological and Artificial Computation: From Neuroscience to Technology. IWANN 1997 (Lecture Notes in Computer Science, 1240)*, Berlín-Heidelberg, Springer, 1997, disponible en **https://doi.org/10.1007/BFb0032504** [consultado el 29 de agosto de 2025].

FREITAG, Markus, y Yaser AL-ONAIZAN, «Beam Search Strategies for Neural Machine Translation», en *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver: Association for Computational Linguistics, 2017, pp. 56–60.

Hansard: Official Proceedings of the House of Commons of Canada, Hull (Canadá), Canadian Government Printing Bureau, 1974-1978.

JIAO, Wenxiang, Wenxuan WANG, Jen-Tse HUANG, Xin XIE, Shuming SHI, y Zhaopeng TU, «Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine»,

arXiv e-prints, Art. no. arXiv:2301.08745, 2013, disponible en **doi:10.48550/arXiv.2301.08745** [consultado el 29 de agosto de 2025.]

KOEHN, Philipp, *et al*., «Statistical Phrase-Based Translation», en *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 127–133.

– *et al*., «Moses: Open Source Toolkit for Statistical Machine Translation», en *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Praga, Association for Computational Linguistics, 2007, pp. 177-180.

– y Rebecca KNOWLES, «Six Challenges for Neural Machine Translation», en *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, Association for Computational Linguistics, 2017, pp. 28-29.

LONGPRE, Shayne *et al*., «The Flan Collection: Designing Data and Methods for Effective Instruction Tuning», en *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, *PMLR*, 202 (2023): 22631–22648, disponible en **https://arxiv.org /abs/2301.13688** [consultado el 29 de agosto de 2025].

MITCHELL, Tim, *Machine Learning*, Columbus (Ohio), McGraw-Hill Science/Engineering/Math, 1997, disponible en **https://www.cs.cmu.edu /~tom/files/MachineLearningTomMitchell.pdf** [consultado el 29 de agosto de 2025.]

MOSLEM, Yasmin, *et al*., «Adaptive Machine Translation with Large Language Models», en *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere, European Association for Machine Translation 2023, pp. 227-237.

OUYANG, Long *et al*., «Training language models to follow instructions with human feedback», en *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook (Nueva York), Curran Associates, 2022, pp. 27730–27744, disponible en **https://arxiv.org/abs/2203.02155** [consultado el 29 de agosto de 2025.]

RAHEJA, Vipul, *et al*., «CoEdIT: Text Editing by Task-Specific Instruction Tuning», en *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapur, Association for Computational Linguistics, 2023, pp. 5274–5291.

ROBINSON Nathaniel, *et al*., «ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages», en *Proceedings of the Eighth Conference on Machine Translation*, Singapur, Association for Computational Linguistics, 2023, pp. 392-418.

SENNRICH, Rico, *et al*., «Neural Machine Translation of Rare Words with Subword Units», en Katrin Erk y Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers, Berlín, Association for Computational Linguistics, 2016, pp. 1715–1725.

SHUMAILOV, Ilia, *et al*., «AI models collapse when trained on recursively generated data», *Nature*, 631 (2024), 755-759.

SINGH, Shivalika, *et al*., «Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning», Lun-Wei Ku, Andre Martins y Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers, Bangkok, Association for Computational Linguistics, 2024, pp. 11521–11567, disponible en **https://aclanthology.org/2024.acl-long.620/** [consultado el 29 de agosto de 2025].

SUTSKEVER, Ilya, *et al*., «Sequence to sequence learning with neural networks», en Zoubin Ghaharamani, Max Welling y Corinna Cortes (eds.), NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 2, Cambridge (Massachussetts), MIT Press, 2014, pp. 3104–3112.

TUROVSKY, Barak, «Ten years of Google Translate", Google Translate Blog, 28 abril 2016.

VASWANI, Ashish, *et al*., «Attention is All you Need», en Isabelle Guyon *et al*. (eds.), *Advances in Neural Processing Systems (NIPS 2017)*, vol. 30 (2017), pp. 5999-6009, disponible en **https://arxiv.org/abs/1706.03762** [consultado el 29 de agosto de 2025].