UAB
Universitat Autònoma de Barcelona

TRADUCCIÓN AUTOMÁTICA NEURONAL PARA LENGUAS CON RECURSOS REDUCIDOS
La evaluación por usuarios según el principio de no inferioridad

María do Campo Bayón

TRADUCCIÓN AUTOMÁTICA NEURONAL PARA LENGUAS CON RECURSOS REDUCIDOS

La evaluación por usuarios según el principio de no inferioridad

María do Campo Bayón

Dirección:
Dra. Pilar Sánchez Gijón

Tesis doctoral
Doctorado en Traducció i Estudis Interculturals
Facultat de Traducció i Interpretació
2023

UAB
Universitat Autònoma de Barcelona

In the end, these pieces show that the concept of minority is worth exploring because it inspires innovation in translation and research.

Lawrence Venuti.

Resumo

A presente tese de investigación ten como obxectivo desenvolver un motor de tradución automática neuronal (TAN) para a combinación castelán-galego e o ámbito das redes sociais. A investigación parte do análise dos procesos habituais de adestramento da TAN dentro da industria da tradución. A partir dese análise, fundamentaranse os pilares teóricos deste traballo. En primeiro lugar, abordarase o multilingüismo e a súa relación coa comunicación dixital en mensaxes curtas, onde se afondarán na súa caracterización e analizaranse os principais desafíos que a tradución automática presenta neste xénero textual. Posteriormente, definirase a tradución automática neuronal, describirase o marco legal e os procedementos habituais para a creación do corpus de adestramento e, finalmente, analizaranse as principais métricas de avaliación da calidade da TAN.

Unha vez descrito o marco teórico e os antecedentes, dentro do marco metodolóxico, describiranse en detalle os procedementos levados a cabo para adestrar o motor de tradución e crear o corpus de adestramento necesario para o seu funcionamento óptimo. Posteriormente, presentarase a estratexia de avaliación da calidade deseñada especificamente para este motor e contexto particular. Este enfoque novidoso incorpora tres métricas de avaliación distintas: BLEU, MQM-DQF e un análise de non inferioridade, co propósito de obter datos cuantitativos e cualitativos exhaustivos sobre as traducións de chíos. A avaliación de non inferioridade, en particular, presenta como unha aproximación innovadora no campo da avaliación da calidade da TAN. Para validar tanto o motor como o instrumento de análise, realízase unha proba piloto inicial. Os resultados e datos obtidos nesta fase piloto empréganse para mellorar o motor e ampliar o corpus de adestramento. Posteriormente, procedese a unha avaliación máis exhaustiva do motor, integrando os datos das tres métricas de avaliación mencionadas anteriormente. A triangulación de resultados proporciona unha avaliación completa da calidade final do motor.

Finalmente, a partir da análise dos datos recollidos, alcanzaranse os obxectivos propostos e extraeranse conclusións sólidas que apoian as hipóteses de partida e contribúen ao coñecemento no campo da TAN e a súa aplicación en contextos específicos de comunicación dixital.

Palabras clave: *tradución automática, lingua minorizada, lingua con recursos reducidos, tradución automática neuronal, avaliación da tradución automática, errores de tradución automática, avaliación de non inferioridade*

Resumen

La presente tesis de investigación tiene como objetivo desarrollar un motor de traducción automática neuronal (TAN) para la combinación castellano-gallego y el ámbito de las redes sociales. La investigación parte del análisis de los procesos habituales de entrenamiento de la TAN dentro de la industria de la traducción. A partir de este análisis, se fundamentarán los pilares teóricos de este trabajo. En primer lugar, se abordará el multilingüismo y su relación con la comunicación digital en mensajes cortos, donde se profundizará en su caracterización y se analizarán los principales desafíos que plantea la traducción automática en este género textual. Posteriormente, se definirá la TAN, se describirá el marco legal y los procedimientos habituales para la creación del corpus de entrenamiento y, finalmente, se analizarán las principales métricas de evaluación de la calidad de la TAN. Una vez descrito el marco teórico y los antecedentes, dentro del marco metodológico, se describen en detalle los procedimientos llevados a cabo para entrenar el motor de traducción y crear el corpus de entrenamiento necesario para su funcionamiento óptimo. Posteriormente, se presenta la estrategia de evaluación de la calidad diseñada específicamente para este motor y contexto particular. Este enfoque novedoso incorpora tres métricas de evaluación distintas: BLEU, MQM-DQF y un análisis de no inferioridad, con el propósito de obtener datos cuantitativos y cualitativos exhaustivos sobre las traducciones de tuits. La evaluación de no inferioridad, en particular, se presenta como una aproximación innovadora en el campo de la evaluación de la calidad de la TAN. Para validar tanto el motor como el instrumento de análisis, se realiza una prueba piloto inicial. Los resultados y datos obtenidos en esta fase piloto se emplean para mejorar el motor y ampliar el corpus de entrenamiento. Posteriormente, se procede a una evaluación más exhaustiva del motor, integrando los datos de las tres métricas de evaluación mencionadas anteriormente. La triangulación de resultados proporciona una evaluación completa de la calidad final del motor. Finalmente, a partir del análisis de los datos recopilados, se alcanzan los objetivos planteados y se extraen conclusiones sólidas que respaldan las hipótesis de partida y contribuyen al conocimiento en el campo de la TAN y su aplicación en contextos específicos de comunicación digital.

Palabras clave: *traducción automática, lengua minorizada, lengua con recursos reducidos, traducción automática neuronal, evaluación de la traducción automática, errores de traducción automática, evaluación de no inferioridad*

Abstract

The present research thesis aims to develop a neural machine translation (NMT) engine for the Spanish-Galician language pair within the domain of social media communication. The research commences with an analysis of the typical NMT training processes employed within the translation industry, laying the foundational theoretical framework for this work. Firstly, the thesis delves into the realm of multilingualism and its relationship with digital communication through short messages, providing an in-depth characterization and addressing the primary challenges posed by automatic translation within this textual genre. Subsequently, it defines neural machine translation, describes the legal framework, and outlines the standard procedures for creating the training corpus. Additionally, it scrutinizes the principal metrics used to assess the quality of NMT.

Once the theoretical framework and antecedents have been elucidated, the methodological section offers a comprehensive description of the procedures used to train the translation engine and construct the necessary training corpus for its optimal functioning. Furthermore, it introduces the quality assessment strategy designed specifically for this engine and context. This innovative approach incorporates three distinct evaluation metrics: BLEU, MQM-DQF, and a non-inferiority analysis, with the aim of obtaining comprehensive quantitative and qualitative data on tweet translations. The non-inferiority assessment emerges as a pioneering approach in the field of NMT quality assessment. To validate both the engine and the analytical instrument, an initial pilot test is conducted. The results and data acquired in this pilot phase are utilized to enhance the engine and expand the training corpus. Subsequently, a more exhaustive evaluation of the engine is performed, integrating the data from the three evaluation metrics. The triangulation of results yields a comprehensive assessment of the final quality of the engine.

Finally, based on the analysis of the collected data, the established objectives are met, and robust conclusions are drawn that support the initial hypotheses. This research contributes to the field of NMT and its application in specific contexts of digital communication.

***Keywords**: machine translation, minority language, less-resourced language, neural machine translation, MT evaluation, MT errors, no-inferiority evaluation*

Resum

La present tesi de recerca té com a objectiu desenvolupar un motor de traducció automàtica neuronal (TAN) per a la combinació de castellà i gallec, amb un enfocament especial en la comunicació digital a través de les xarxes socials. La recerca parteix de l'anàlisi dels processos habituals d'entrenament de la TAN dins de la indústria de la traducció. A partir d'aquesta anàlisi, es fonamentaran els fonaments teòrics d'aquest treball. En primer lloc, s'abordarà el multilingüisme i la seva relació amb la comunicació digital en missatges curts, on s'aprofundirà en la seva caracterització i s'analitzaran els principals reptes que planteja la traducció automàtica en aquest gènere textual específic. Posteriorment, es definirà la traducció automàtica neuronal, es descriurà el marc legal i els procediments habituals per a la creació del corpus d'entrenament i, finalment, s'analitzaran les principals mètriques d'avaluació de la qualitat de la TAN.

Un cop descrit el marc teòric i els antecedents, dins del marc metodològic, es descriuen en detall els procediments duts a terme per entrenar el motor de traducció i crear el corpus d'entrenament necessari per al seu funcionament òptim. Posteriorment, es presenta l'estratègia d'avaluació de la qualitat dissenyada específicament per a aquest motor i context particular. Aquest enfocament innovador incorpora tres mètriques d'avaluació diferents: BLEU, MQM-DQF i una anàlisi de no inferioritat, amb la finalitat d'obtenir dades quantitatives i qualitatives exhaustives sobre les traduccions de tuïts. L'avaluació de no inferioritat, en particular, es presenta com una aproximació innovadora en el camp de l'avaluació de la qualitat de la TAN. Per validar tant el motor com l'instrument d'anàlisi, es realitza una prova pilot inicial. Els resultats i les dades obtinguts en aquesta fase pilot s'empren per millorar el motor i ampliar el corpus d'entrenament. Posteriorment, es procedeix a una avaluació més exhaustiva del motor, integrant les dades de les tres mètriques d'avaluació esmentades anteriorment. La triangulació de resultats proporciona una avaluació completa de la qualitat final del motor.

Finalment, a partir de l'anàlisi de les dades recopilades, s'assoleixen els objectius plantejats i s'extrauen conclusions sòlides que recolzen les hipòtesis de partida i contribueixen al coneixement en el camp de la TAN i la seva aplicació en contextos específics de comunicació digital.

Paraules clau: *traducció automàtica, llengua minoritzada, llengua amb recursos reduïts, traducció automàtica neuronal, avaluació de la traducció automàtica, errors de traducció automàtica, avaluació de no inferioritat*

LISTA DE ABREVIATURAS

BLEU	*BiLingual Evaluation Understudy*
DQF	*Dynamic Quality Framework*
ES	*Castellano*
GL	*Gallego*
LLM	*Large Language Models*
LP	*Lengua de partida*
LD	*Lengua de destino*
LSP	*Language Service Provider*
MQM	*Multidimensional Quality Metrics*
MT	*Machine Translation*
NMT	*Neural Machine Translation*
TA	*Traducción Automática*
TAO	*Traducción Asistida por Ordenador*
TAN	*Traducción Automática Neuronal*

Índice de contenidos

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

INTRODUCCIÓN

En los últimos años, hemos asistido al asentamiento de los motores neuronales en el mercado de la traducción como tecnología prioritaria. Sin embargo, no todas las combinaciones de idiomas tienen un rendimiento satisfactorio, especialmente si hablamos de lenguas con recursos reducidos. Este es el caso particular de mi lengua materna, el gallego. En una investigación anterior (do Campo *et al.*, 2019) se demostró que la tecnología neuronal no era todavía eficaz en la combinación castellano-gallego. Consecuentemente, me propongo estudiar cómo actualizar la tecnología de traducción automática existente en gallego a la era neuronal mediante la creación de un motor neuronal propio, sin perder de vista la dimensión de dicha lengua como lengua con recursos reducidos. Por ello, me interesaré por los recursos textuales disponibles, por la optimización de las estrategias de ampliación de datos, por el proceso de entrenamiento y por las iniciativas realizadas hasta el momento. Así, el objetivo principal se mantiene en el prisma de la relación dependiente entre traducción automática y lenguas con pocos recursos. Pretendemos, pues, hacer una investigación útil para lenguas con pocos recursos, que en muchas ocasiones son también lenguas de uso minoritario o minorizado, pero que, dado que no es necesario ser una lengua minorizada para tener pocos recursos, nos referiremos a ellas siempre como lenguas con pocos recursos.

Como hablante de la lengua gallega, me parece de especial interés también emplear el motor en un tipo de texto determinado, los textos propios de las redes sociales, con el objetivo de promover el uso y de presencia de la lengua en Internet, especialmente entre la población más joven. Dado que la pérdida de hablantes de gallego se lleva produciendo en los extractos más jóvenes de la población (IGE, 2018), queremos dotar al motor de un uso social de fomento de la lengua y poner un grano de arena en la lucha para incrementar el uso de la lengua.

Por otro lado, para poder evaluar el futuro motor, nos acercamos a la evaluación humana desde la premisa de que las traducciones de un mismo texto pueden diferir y ser totalmente correctas al mismo tiempo, por lo que es necesario un nuevo procedimiento de evaluación que no tome exclusivamente como referencia una traducción *gold standard*.

En resumen, debido a que el gallego como lengua minorizada tiene unas particulares especiales, este trabajo se asienta en la necesidad de abrir una nueva línea de investigación sobre entrenamiento de motores neuronales de traducción automática en lenguas con recursos lingüísticos reducidos, sobre las estrategias que se pueden emplear para optimizar los pocos recursos disponibles y sobre un nuevo procedimiento de evaluación.

1. Objetivos

El objetivo general de esta tesis doctoral es desarrollar un motor de traducción automática neuronal (TAN) de alta calidad y que permita su uso sin la intervención de poseditores para la combinación castellano-gallego y el ámbito de las redes sociales. Para alcanzar este objetivo principal se establecerán los siguientes objetivos específicos:

- Profundizar en las prácticas comunes en la industria de la traducción con respecto al entrenamiento de motores neuronales de traducción automática. Para ello, se establecen otros objetivos concretos. Por un lado, consultar las fuentes secundarias y, por otro lado, llevar a cabo un estudio de campo.
- Entrenar un motor de alta calidad para el género textual propio de las redes sociales y, en concreto, de Twitter. De nuevo, dentro de este objetivo, se debe primero investigar las estrategias para la creación y optimización de datos necesarios para el entrenamiento de motores neuronales, investigar el proceso de entrenamiento de motores neuronales y seleccionar la solución tecnológica más apropiada.
- Generar un corpus suficiente en calidad y cantidad para el entrenamiento del motor. Dentro de este objetivo, primero se debe recopilar los recursos disponibles para el entrenamiento de motores neuronales en la combinación de trabajo y, si no son suficientes, generar nuestro corpus propio. Por último, utilizar los datos del modo más conveniente para garantizar la calidad del motor
- Diseñar una metodología de evaluación conveniente del motor neuronal entrenado dentro del contexto de las redes sociales. Para ello, primero se debe utilizar los métodos estándar y evaluar si resultan suficientes. Finalmente, proponer un método apropiado para la evaluación de texto de redacción libre con TAN de alta calidad.

Consideramos que la consecución de estos objetivos específicos, algunos de ellos dependientes entre sí, nos permitirá alcanzar el objetivo principal fijado.

2.Hipótesis de trabajo

La hipótesis de investigación subyacente sobre la que gira esta investigación se concreta del siguiente modo: la tecnología actual y el gran volumen de información disponible en formato digital, incluso en el caso de una lengua con pocos recursos como el gallego, permite generar motores de TAN de alta calidad para contextos de uso específicos de redacción libre, como las redes sociales y, en concreto, tuits. Las siguientes preguntas específicas ayudarán a refutar la hipótesis principal:

- ¿Qué cantidad de datos establecen las empresas de traducción como mínima para un entrenamiento satisfactorio?
- ¿Qué categoría de corpus y qué cantidad de colecciones en forma de corpus están disponibles en la combinación de trabajo?
- ¿Cuáles son las mejores estrategias de optimización de recursos en la combinación de lenguas?
- ¿Es posible recopilar los datos suficientes para optimizar un motor?
- ¿Cuál es el método más oportuno para evaluar el rendimiento del motor dentro del contexto de las redes sociales y las lenguas con pocos recursos?
- ¿Cuáles son los puntos fuertes y los puntos débiles del motor entrenado?
- ¿Los resultados del motor en términos de calidad, varían en función de la longitud del texto que se traduce?

Estas preguntas nos permiten identificar varias subhipótesis o hipótesis parciales. La primera consiste en que la TAN para las redes sociales en un contexto de lenguas con pocos recursos necesitará de una metodología de evaluación diferente a los métodos de evaluación actuales. En segundo lugar, dada la naturaleza del género que vamos a estudiar, tuit, prevemos también que el rendimiento del motor variará según la longitud de los textos evaluados.

3.Estructura de la tesis

La estructura de esta tesis se desarrolla de manera minuciosa y detallada, abarcando diversos aspectos cruciales en el ámbito de la TAN y la comunicación digital en mensajes cortos. A lo largo de las siguientes páginas, exploraremos en profundidad cada uno de los elementos que componen esta investigación.

Primero se presentan el marco teórico, los antecedentes y la contextualización que sustentan esta investigación. En este grupo de capítulos se presentan no solo los conceptos fundamentales relacionados con la TAN y la comunicación digital, sino también los antecedentes que contextualizan las prácticas y procesos en TAN en la industria de la traducción. Dentro de este marco teórico, se destacan las principales herramientas y técnicas que son esenciales para el desarrollo de sistemas de TAN en el contexto de la comunicación digital en mensajes cortos. Esto incluye la creación de corpus multilingües, un proceso crucial para garantizar la calidad de las traducciones automáticas. También se aborda el preprocesamiento y posprocesamiento de las traducciones automáticas, dos etapas cruciales que influyen directamente en la precisión y fluidez de los resultados. Finalmente, se exploran en detalle los métodos de evaluación de la calidad de la traducción, un aspecto fundamental para medir el desempeño de los sistemas de traducción automática.

Después de sentar las bases teóricas y proporcionar una fotografía de la industria, nos sumergiremos en el segundo bloque de capítulos, que constituye el marco metodológico de esta tesis. En esta sección, se describe minuciosamente el proceso de creación del motor de traducción automática neuronal, específicamente diseñado para traducir del castellano al gallego en el contexto de las redes sociales. Este proceso incluye la selección de la tecnología neuronal adecuada, así como la recopilación y preparación de datos específicos para este proyecto. Además, se aborda el procedimiento de evaluación diseñado para medir la calidad del motor entrenado.

Finalmente, encontraremos la sección de conclusiones finales y perspectivas de trabajo en el futuro. Aquí, se resumen las principales contribuciones de esta investigación y se destacan los hallazgos más significativos. Además, se abre una puerta a la reflexión sobre las posibles direcciones futuras que podrían llevar a mejoras y avances adicionales en el campo de la traducción automática.

MARCO TEÓRICO, ANTECEDENTES Y CONTEXTUALIZACIÓN

La comunicación digital ha sido uno de los campos que más ha evolucionado en las últimas décadas, lo que ha llevado a una mayor interacción entre personas de diferentes países y culturas. En este contexto, el multilingüismo se ha convertido en un aspecto clave para la comunicación global.

En la era contemporánea de la comunicación digital, el multilingüismo se erige como una fuerza trascendental que enriquece la interacción global. La diversidad lingüística se manifiesta en plataformas digitales, donde la coexistencia de múltiples idiomas genera una red de conexiones culturales y facilita el intercambio de ideas entre personas de distintos rincones del mundo. Sin embargo, este fenómeno también plantea desafíos intrigantes, como la necesidad de herramientas de traducción eficientes y la preservación de la autenticidad en la expresión lingüística. En este contexto, el multilingüismo no solo trasciende las barreras idiomáticas, sino que redefine la forma en que forjamos vínculos y compartimos conocimiento en la aldea global digital. Por tanto, en este bloque de capítulos se contextualizará en primer lugar los procedimientos que siguen las empresas de cara a la traducción automática con la ayuda de una encuesta realizada a empresas miembro de la asociación TAUS. Posteriormente, gracias al análisis de estas respuestas, se fundamentarán los pilares teóricos de esta investigación. En primer lugar, se abordará el multilingüismo y su relación con la comunicación digital en mensajes cortos, donde se profundizará en su caracterización y se analizarán los principales desafíos que plantea la traducción automática en este género textual.

En el siguiente capítulo, se definirá la traducción automática y se especificarán las tecnologías y procedimientos más empleados. Posteriormente, se analizará el marco legal para la creación de corpus multilingües, que es esencial para la obtención de datos de alta calidad y la mejora de los sistemas de traducción automática. Asimismo, se presentarán las herramientas para la creación de corpus sintéticos y los recursos para el preprocesamiento y posprocesamiento de la traducción automática.

Por último, se analizarán los instrumentos de evaluación métrica, que son esenciales para la evaluación de la calidad de la traducción automática. Se examinarán las principales métricas utilizadas en la evaluación de la calidad de la traducción automática, así como las limitaciones de estas métricas.

En resumen, en este bloque de capítulos se realizará una fotografía de la industria y se presentarán las principales herramientas y técnicas para el desarrollo de sistemas de traducción automática en el ámbito de la comunicación digital en mensajes cortos, incluyendo la creación de corpus multilingües, el preprocesamiento y posprocesamiento de la traducción automática y la evaluación de la calidad de la traducción.

CAPÍTULO I: Fotografía de los procesos empleados en traducción automática dentro la industria de la traducción

1. Introducción

Según el informe ELIS del año 2023 sobre las últimas tendencias en el mercado de la traducción, la Traducción Automática (TA) vuelve a revalidar su posición como uno de los servicios con mayor crecimiento exponencial (ELIS 2023: 25), lo que la posiciona como un recurso más dentro de la industria de la traducción, al mismo nivel que las memorias de traducción (TM) y los glosarios o equivalente. No sorprende, por tanto, que los datos de 2023 confirmen que los proveedores de servicios lingüísticos (LSP, por sus siglas en inglés) optan por invertir en tecnología, y dentro de esta categoría, principalmente en traducción automática (ELIS 2023: 38). Podemos afirmar que el uso de la TA es un hecho, aunque su implantación en empresas y su desarrollo sea un camino poco explorado en lenguas minoritarias. Por esta razón, el primer capítulo de esta investigación se centrará en realizar una fotografía de cuáles son los procesos empleados en las LSP y qué recursos utilizan para, a partir de los resultados obtenidos, estructurar el marco teórico. La idea detrás de esta primera indagación es marcar un punto de inicio sobre el que proponer un protocolo común para lenguas con recursos reducidos y determinar unos escenarios mínimos que puedan ser extrapolables, al menos, a lenguas románicas.

La bibliografía consultada arroja algo de luz acerca de las motivaciones de una empresa para invertir en TA y refleja los cambios tecnológicos de los últimos años sobre la TA y las diferentes tendencias en función de las lenguas de trabajo. Habitualmente, los tres factores que más se destacan son: pedido expreso de los clientes, tarifas de traducción y plazos (Torres-Hostench *et al.*, 2016). Sin embargo, existe poca información sobre cómo y cuándo una empresa decide incluir la TA en su flujo de producción. Consecuentemente, decidimos recoger la posición de algunas empresas representativas a través de una encuesta. El cuestionario se realizó durante el verano de 2020 y se envió a empresas miembro de la organización TAUS con la intención de determinar qué es lo que motivaba a estas empresas

participantes a invertir en TA y qué procesos seguían para determinar que un nuevo motor de TA era apto para usarlo en un entorno de producción.

2. La encuesta

Como ya hemos mencionado más arriba, la encuesta estaba dirigida a proveedores de servicios lingüísticos y el objetivo principal era saber qué cantidad de datos necesitaban las LSP para crear y entrenar motores de TA. Para ello, se diseñó un cuestionario en inglés como lengua vehicular de 17 preguntas para obtener datos cuantitativos y cualitativos de distinta naturaleza. Esta información se puede dividir en dos grandes grupos: información descriptiva sobre la LSP recogida en preguntas de respuesta múltiple e información sobre el uso y/o desarrollo de TA en formato de respuesta corta o larga para evitar guiar al encuestado, ya que no existe un estándar al respecto. De esta manera, se pretendía obtener información sobre el perfil de la empresa, las fases de entrenamiento, las fases de pruebas y los procesos empleados para evaluar la calidad.

Las preguntas y los umbrales se fijaron en colaboración con un experto en TA, Diego Bartolomé. Esto permitió crear un instrumento de recogida de datos preciso y adecuado. Fruto de esta colaboración, surgió la idea de plantear la recogida de esta información cuantitativa mediante otro instrumento como la encuesta anónima, y de buscar la colaboración de un organismo internacional como TAUS o GALA para conseguir una mayor participación y evitar la negativa de muchas empresas a proporcionar información confidencial. Finalmente, conseguimos llegar a un acuerdo con TAUS y lanzar la encuesta dentro de sus campañas de fidelización y creación de comunidad.

La encuesta estuvo activa durante todo el mes de septiembre de 2020 a través de *Google Forms*. Contamos con la participación de 9 empresas. A pesar de que no es una participación alta, sí fue representativa de la tipología de empresas para poder extraer ciertas conclusiones a partir de las cuales estructurar el marco teórico, principalmente el tipo de datos empleados y la cantidad mínima de datos para poder entrenar un motor, el procedimiento de evaluación de la TA y el proceso de entrenamiento.

En el Anexo I, se puede consultar la encuesta completa.

A continuación, en el apartado 3 presentamos las preguntas que permiten describir el perfil de los participantes junto con las respuestas obtenidas. Por su parte, la sección 4, recoge las respuestas en relación con la implementación de la TA.

3.Perfil de los participantes

La primera sección de la encuesta incluye preguntas relacionadas con la actividad de las LSP para obtener información sobre su perfil y también para poder realizar un análisis detallado de los datos obtenidos. Se les solicita información sobre el tamaño de la empresa en términos de plantilla, la ubicación de sus oficinas principales y si tienen o no un departamento específico de TA.

El 55,6 % de los participantes representan a LSP de entre 26 y 50 empleados y el 44,4 % restante empresas con más de 100 empleados.

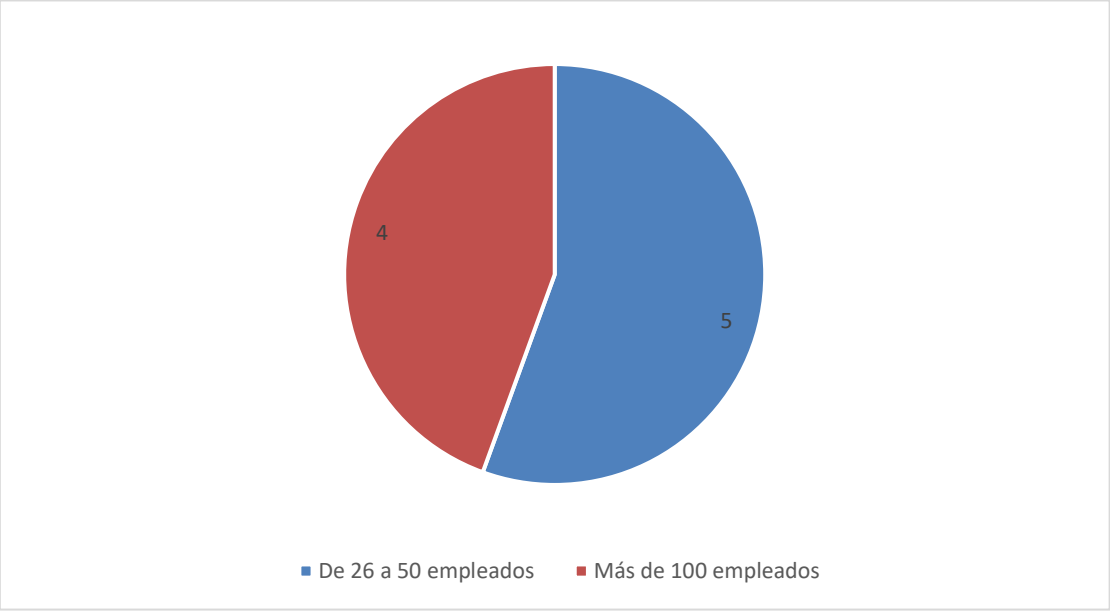


Ilustración 1. Tamaño de las LSP en términos de plantilla

En cuanto a las oficinas principales los encuestados podían seleccionar más de una ubicación. Con todo, la mayoría están situadas en Europa y América, y solamente el 8 % en Asia.

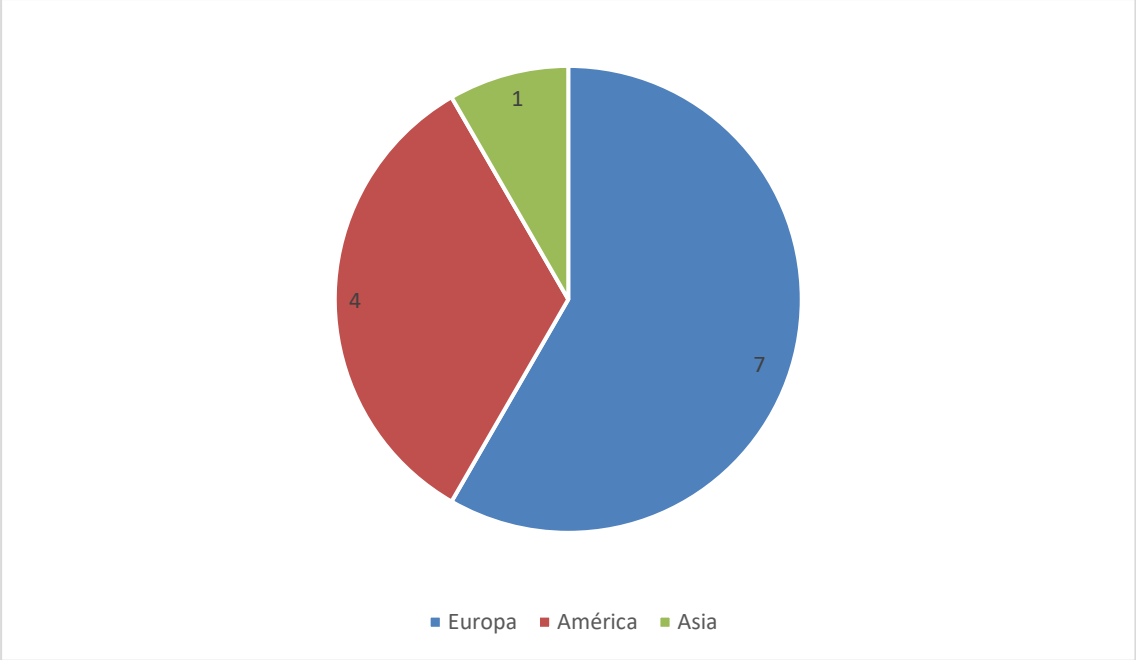


Ilustración 2. Ubicación de las oficinas principales de las LSP

En lo relativo a la pregunta sobre el departamento específico de TA, el 44,4 % de los participantes tienen su propio departamento de TA, mientras que el 33,3 % responde negativamente. Por su parte, un 11,1 % de las empresas responde que tienen empleados dentro de su departamento de producción responsables de todas las tareas relaciones con TA y un mismo porcentaje admitió que casi tienen un departamento de TA.

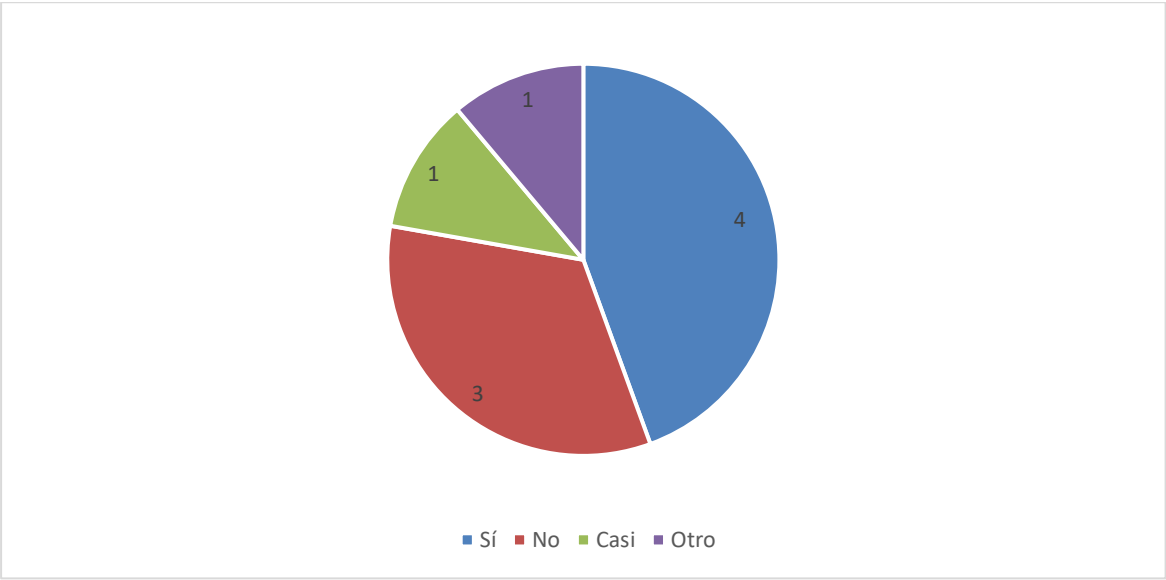


Ilustración 3. Existencia de un departamento propio de TA dentro de las LSP

Si analizamos los resultados, por tanto, podemos concluir que el perfil principal del participante es una empresa de mediana a grande, principalmente asentada en Europa con bien un departamento específico de MT o con una dedicación especial para tareas de TA. Perfil que se corresponde en gran medida con el encuestado por ELIS (2023: 7).

4. Procesos de traducción automática

Dentro de este apartado, se abordan las preguntas relativas a los procesos de TA. En primer lugar, se les pregunta la motivación para invertir en TA. Posteriormente, si han identificado o no un número de datos mínimos de entrenamiento, qué fases de entrenamiento llevan a cabo, qué procedimientos siguen para implementar un motor en producción, qué indicadores de evaluación utilizan y si realizan mejoras de los motores entrenados. Finalmente, se les pregunta si se ven afectados por la COVID-19.

4.1. Motivación para invertir en TA

En primer lugar, queremos determinar qué es lo que motiva a las empresas a invertir en TA. En esta pregunta al participante se le proporciona 4 opciones: cliente, plazo, tarifa de traducción y la combinación de los tres. Si ninguna de las opciones se ajusta a su realidad, los participantes pueden introducir su propia respuesta. En esta pregunta, se puede escoger una o varias opciones.

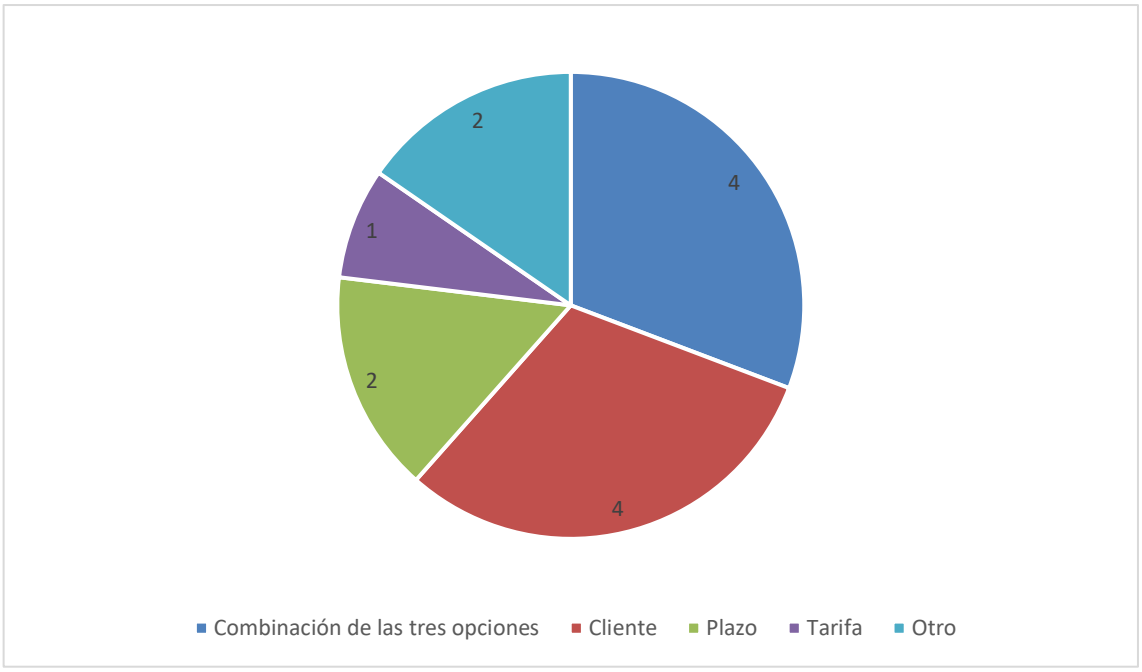


Ilustración 4. Motivos de las LSP para invertir en TA

Un 31 % de los encuestados escoge la cuarta opción: «una combinación de las tres opciones». Sin embargo, también cabe resaltar el mismo porcentaje obtenido debido a las exigencias del cliente. En tercer lugar, los encuestados sitúan al plazo con un 15 % de respuestas. A una distancia considerable de las otras tres opciones, con un 8 %, se sitúan las tarifas de traducción. Dentro del apartado de respuesta libre, dos de las empresas esbozan nuevas razones, aunque más que motivos para invertir, quizás apuntan más a la viabilidad de usar TA: por un lado, el departamento de control de calidad (QA) y, por otro lado, la cantidad de datos propios.

4.2. Datos de entrenamiento

En lo referente a los datos de entrenamiento, la encuesta permite recabar información clave para el desarrollo de esta investigación sobre dos tópicos diferentes: el tipo de recursos empleados y la cantidad mínima de datos necesaria para entrenar los motores de TA.

Por ello, la pregunta que se les hace es: «¿qué tipos de recursos usas para entrenar un motor?». Uno de los principales recursos que usan para entrenar un nuevo motor es la memoria de TA (50 %) seguida de una combinación de todos los recursos (17 %), un corpus específico del género textual (17 %) y un corpus genérico (8 %). Esto nos puede llevar a asumir que la mayoría de las LSP se inclinan más hacia la idea de crear motores especializados de TA en lugar de usar motores genéricos, ya que un 8 % también reconoce usar la terminología propia del cliente.

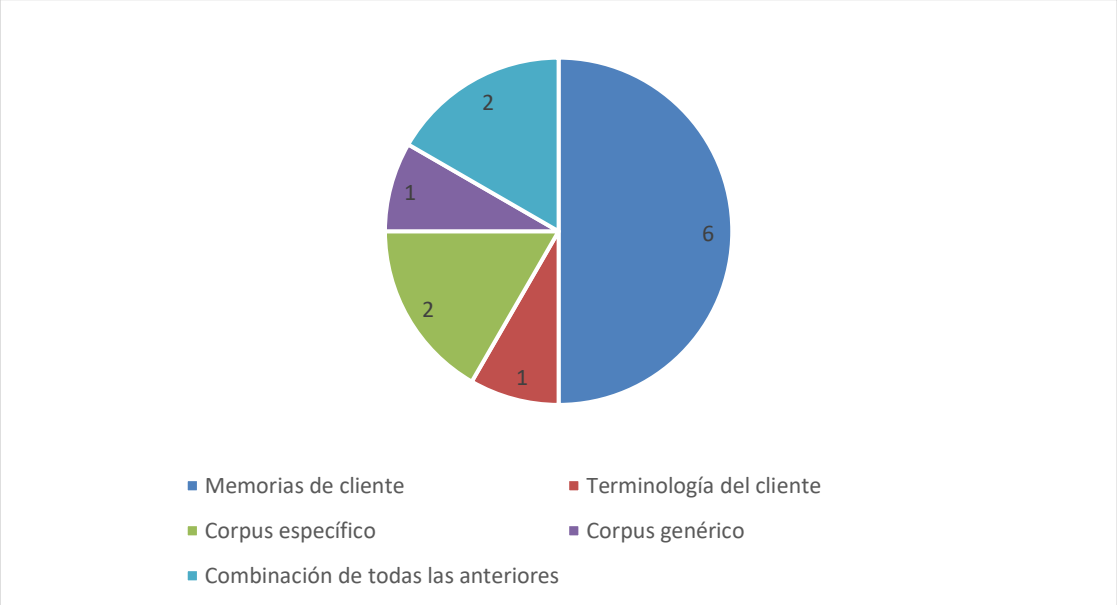


Ilustración 5. Recursos usados para entrenar motores de TA

Una vez que las LSP identifican los recursos principales, la siguiente pregunta es: «¿Has establecido una cantidad mínima de datos necesaria para entrenar un motor de TA?». En este apartado, nos encontramos con una dualidad contrapuesta: el 55,6 % de los encuestados responden que sí han establecido esta cantidad para uno o más pares de lenguas, mientras que el 44,4 % no lo ha establecido. Tras estas respuestas, es evidente la necesidad de más trabajos al respecto, tal y como adelantamos en nuestra introducción.

Si la respuesta es afirmativa, se les pide además que indiquen esa cantidad. De nuevo, nos encontramos con una gran discrepancia, aunque todas las respuestas apuntan a grandes números. El rango va de los 10-15 millones de pares a al menos 80 000 pares de segmentos. Cada par de lenguas y género textual necesitará más o menos datos, pero a grandes rasgos consideramos que se puede establecer un umbral mínimo de ente dos millones a cinco millones de pares de segmento. En consecuencia, buscaremos confirmar esta afirmación en nuestra investigación.

4.3. Fase de entrenamiento

La fase de entrenamiento de los motores de TA, tanto estadísticos como neuronales, es una tarea difícil de delimitar para todas las LSP. Habitualmente, los especialistas en TA tienen que responder a preguntas sobre el tipo y número de evaluaciones, el tipo de métricas y la cantidad de texto evaluado. Tal y como avanzamos en la explicación del diseño de la encuesta, no existe ningún estándar al respecto y cada investigador o empresa tiende a adaptarlo a sus propias necesidades. No esperamos, por tanto, obtener respuestas claras al respecto, pero sí un abanico lo suficientemente amplio como para poder proponer nuestro propio escenario. Por esta razón, el tipo de pregunta que se elige es una pregunta de respuesta corta que les permita responder de forma no guiada. Las respuestas a la pregunta «¿qué procedimiento sigues para evaluar un motor de TA?» revelan que las empresas siguen una mezcla de evaluaciones automáticas y humanas. La mayoría de encuestados combinan métricas automáticas con pruebas en las que se involucra a lingüistas (tareas de posedición, tareas de clasificación manual, etc.). Un tercio de las empresas escoge realizar únicamente evaluaciones humanas como, por ejemplo, la posedición realizada por un traductor nativo, revisión humana o la externalización a especialistas de TA. A pesar de esto, también hay empresas que afirman que emplean exclusivamente métricas automáticas como BLEU, TER e informes de comparación. Otras informan que definieron un proceso específicamente creado para sus tipos de proyectos y clientes. Un ejemplo de esto es el

siguiente proceso empleado por uno de los encuestados: realización de métricas automáticas contra una traducción humana de referencia en un set de pruebas combinadas con una evaluación humana de unas 1000 palabras procedentes de una muestra representativa, realización de métricas automáticas para comparar un motor entrenado contra un motor genérico y, por último, repetición de la fase de entrenamiento y pruebas hasta alcanzar una puntuación aceptable.

En cuanto a los tipos de indicadores usados, se sitúan en sintonía con los procesos seguidos. En la pregunta de respuesta múltiple, «¿qué tipo de indicadores usas?», se le ofrece al participante varias opciones. Las respuestas evidencian que todas las empresas utilizan todos los indicadores disponibles:

- evaluación automática
- evaluación humana
- distancia de edición
- esfuerzo de posedición
- pruebas de productividad (únicamente en proyectos a largo plazo)

Paralelamente, encontramos una mayor divergencia en las comparaciones realizadas. Aquí se les pregunta a las empresas qué comparaciones efectúan y se les da varias opciones. Los encuestados pueden elegir más de una opción.

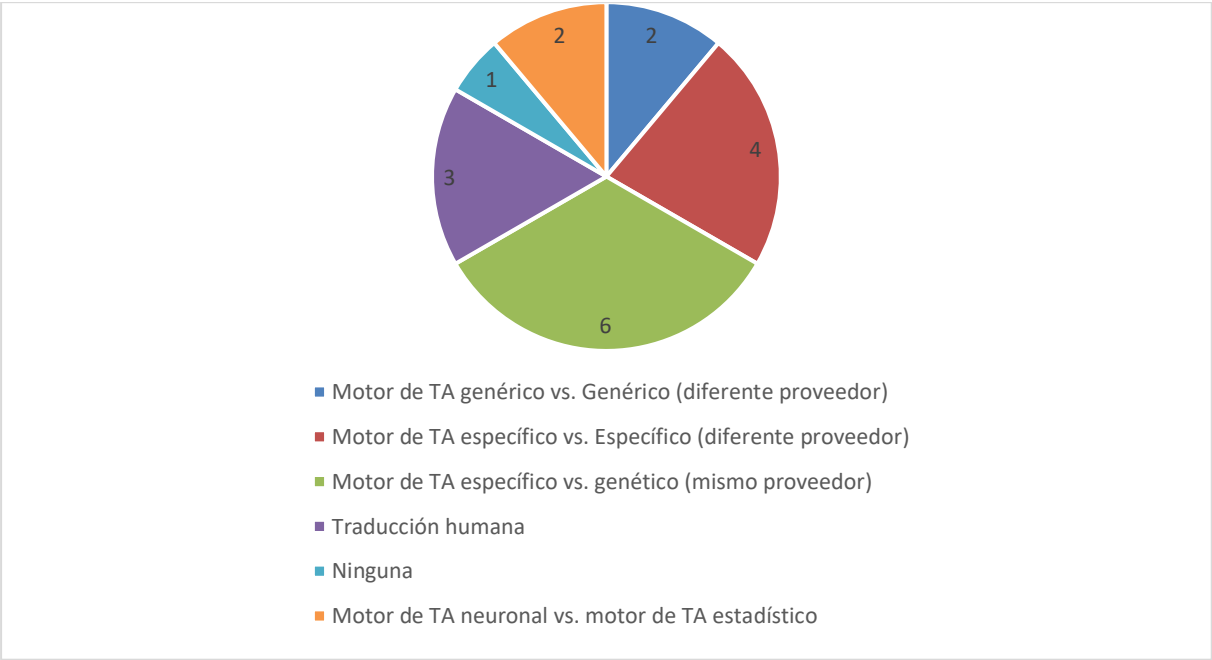


Ilustración 6. Comparaciones de TA

Si observamos las respuestas, vemos que la mayoría de las empresas compara el resultado de un motor de TA específico con el resultado de un motor genérico (mismo tipo de motor), el resultado del motor de TA con una traducción humana de referencia y la traducción sin poseditar de un motor de TA específico contra otra traducción de otro motor específico de otro proveedor. También hubo empresas que reportaron ejecutar comparaciones entre dos motores genéricos diferentes (distinto proveedor) y comparaciones entre motores neuronales y motores estadísticos.

4.4. Fase de adopción o no adopción

Todas las LSP se enfrentan a la misma pregunta en algún momento de su entrenamiento: ¿se puede usar ya mi motor en una fase de producción? Trasladamos esta misma pregunta en formato respuesta libre a todos los encuestados para poder arrojar algo de luz al proceso de toma de decisiones.

Todas las respuestas tienen en común una palabra y esa es la calidad. Sin embargo, la calidad, como veremos en capítulos posteriores, es un término que genera controversia en la teoría de la traducción, por lo que es importante determinar en qué términos las empresas entienden la calidad. El 50 % de los encuestados entienden la calidad en términos de aumento de la productividad, por ejemplo, en los siguientes casos:

- Si la posedición es más rápida que la traducción humana.
- Si la distancia de edición y el plazo se reduce.
- Si el tiempo empleado en realizar la posedición es menor que la traducción de cero.

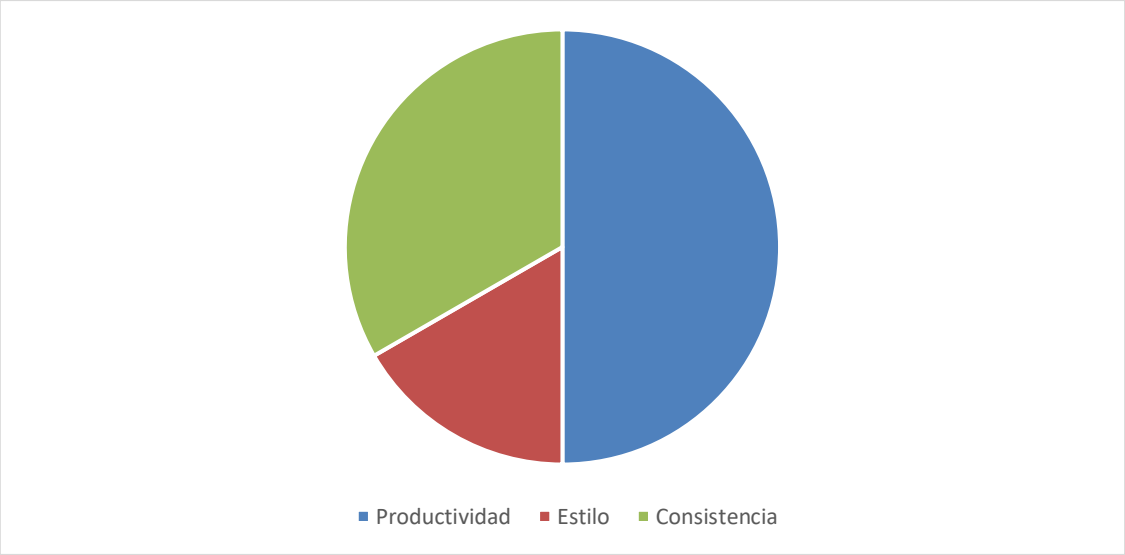


Ilustración 7. Motivos para implementar el motor de TA en los procesos de producción

Otros (17 %) consideran un motor como apto para ser usado si el estilo es aceptable (estilo formal versus estilo informal). Por último, un buen porcentaje de encuestados (33 %) confía en la consistencia de sus ensayos, en si obtienen métricas e índices aceptables y en si obtienen el visto bueno de los evaluadores humanos.

4.5. Indicadores de evaluación aceptables para usar TA

Dado que establecer cuáles son los umbrales de aprobado en los diferentes indicadores de calidad es una tarea titánica y subjetiva, la encuesta plantea dos escenarios para obtener esta información. En primer lugar, pedimos a los encuestados en una pregunta de respuesta libre que nos indiquen sus umbrales mínimos para que una evaluación de TA sea satisfactoria en proyectos de documentación técnica con bajo impacto y/o visibilidad. De entre las 9 LSP, solo 4 realizan este tipo de traducciones. Cada una de las respuestas es diferente por lo que se desglosan las cuatro:

- Buenos índices automáticos.
- Mayor productividad y precio inferior de la posedición frente a la traducción humana.
- < 35 % de distancia de edición y dependiendo de las expectativas de calidad ofrecimiento de posedición completa, posedición parcial o traducción sin poseditar.
- Ningún umbral, ya que depende del presupuesto y del plazo.

Posteriormente, realizamos la misma pregunta, pero esta vez para un proyecto con un impacto y/o visibilidad alto, y las respuestas, de nuevo, difieren bastante por lo que se desglosan todas:

- Aprobación del departamento de QA.
- Mayor productividad y precio inferior de la posedición frente a la traducción humana.
- < 25 % de distancia de edición, ofreciendo posedición complete y, a veces, algún tipo de revisión principalmente de estilo y fluidez.
- 30 HTER.
- Ningún umbral, ya que depende del presupuesto y del plazo.

4.6. Fase de mejoras del motor

Otra información que nos parece interesante recabar es si las empresas siguen mejorando sus motores. Por ello, se les pregunta en formato respuesta libre por qué implementarían mejoras en un motor que ya está en producción.

Ilustración 8. Motivos para entrenar de nuevo un motor de TA

Las LSP afirman que la mejora depende del rendimiento de los poseditores (velocidad y calidad) y de las necesidades del cliente. En menor medida, 11,1 % de las respuestas, la tarifa es otro de los motivos por los que realizar mejoras en los motores.

4.7. Comentarios sobre la traducción

Habitualmente, cuando se entrena y prueba un motor de TA, existe la tendencia a mirar más las métricas, indicadores y evaluaciones sin considerar la perspectiva humana. Por esta razón, decidimos incluir una pregunta específica de respuesta sí/no sobre si las empresas tienen en cuenta los comentarios del poseditor/traductor acerca de la traducción sin poseditar. El 88,9 % de los encuestados afirman que tienen en cuenta los comentarios del traductor, aunque no se pregunta cómo los recogen o procesan.

4.8. Afectaciones de la COVID-19

A pesar de que la información relativa a la pandemia mundial del COVID-19 originada a finales del 2019 no es objeto de investigación de este trabajo, decidimos incluir una pregunta al respecto para dotar de contexto temporal a la encuesta realizada durante el verano de 2020. Por ello, preguntamos si la pandemia está teniendo o no un impacto en el

lanzamiento de estrategias específicas de TA. Diferentes empresas coinciden en que la COVID-19 no está afectando a sus decisiones sobre la inversión en TA, mientras que un número algo menor de respuesta indican lo contrario. También se constata que se está realizando un mayor esfuerzo en investigar datos médicos para ayudar a la sociedad. Por último, no falta quien no está seguro de sus implicaciones.

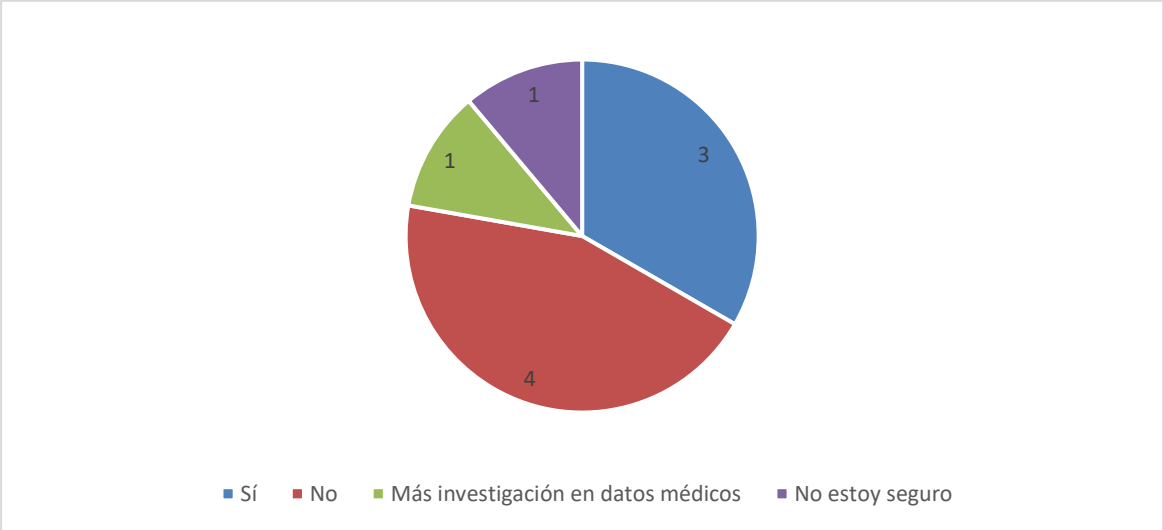


Ilustración 9. Afectación del COVID-19 a las LSP

Para aquellos que afirman verse afectados por la COVID-19, también les pedimos que nos faciliten el por qué en formato respuesta libre. La mayoría parece estar de acuerdo en que está provocando el aumento del uso en TA y otra parte más pequeña en que está provocando el trabajo en remoto.

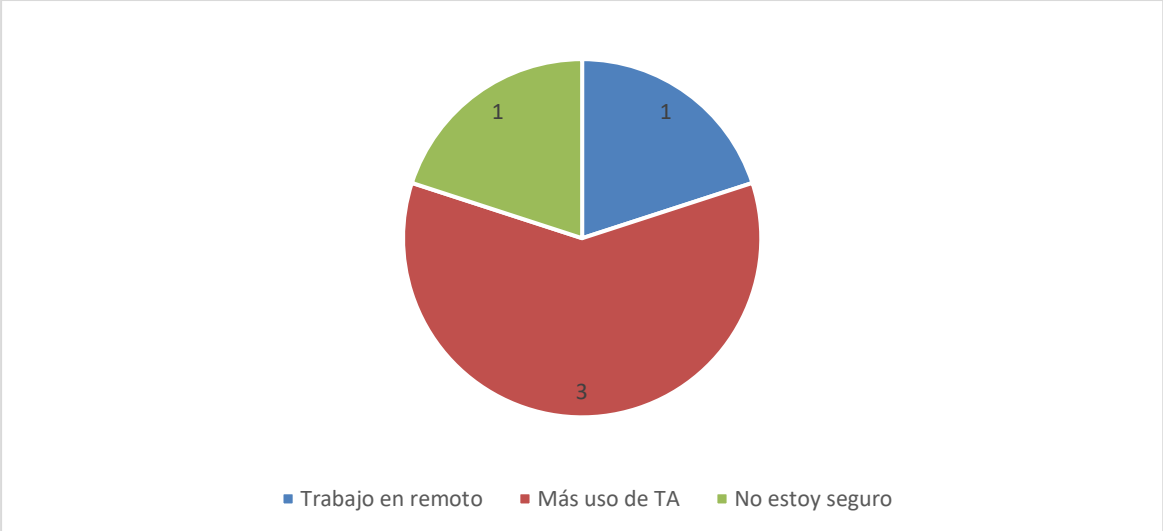


Ilustración 10. ¿Cómo está afectando la COVID-19 a las LSP?

Al analizar cualitativamente las respuestas sobre el aumento de la TA, encontramos dos argumentos principales: la urgencia de algunos proyectos y el mayor interés de los clientes en los servicios de traducción automática y posedición.

Ilustración 11. Razones del aumento del uso de la TA

Si comparamos estos datos de 2020 con la última encuesta ELIS de 2023, vemos que esta tendencia se consolida, ya que la inteligencia artificial y especialmente la traducción automática sigue siendo la tecnología en la más han invertido y prevén invertir las empresas.

5. Consideraciones finales

Esta primera visión introductoria a los procesos de la industria de la traducción de cara a la traducción automática nos ha permitido plantear la hoja de ruta de esta investigación. Si bien es cierto que el número de respuestas obtenidas no permite representar todas las prácticas del sector, no es menos cierto que todas ellas son pertinentes y que provienen de empresas que suelen marcar la tendencia que se seguirá en el futuro más inmediato. A esto hay que sumarle también el carácter reservado de muchas empresas de traducción en facilitarnos información sobre sus procesos y datos empleados.

Tal y como anticipábamos en la introducción, no existe un consenso en la cantidad mínima de datos que se necesita para entrenar un motor de TA ni tampoco un procedimiento estándar para evaluar su calidad. Si bien, gracias a las respuestas de estas empresas, sí podemos afirmar que al menos se necesita llegar a una cifra de millones para empezar a

probar el rendimiento. Con nuestra investigación, buscaremos, por tanto, confirmar ese umbral mínimo en nuestras lenguas de trabajo. De la misma forma, al existir diversas métricas de evaluación de la TA, combinadas también de distinta forma, pretendemos diseñar y validar un procedimiento de evaluación de la calidad de la TA que se pueda replicar, en cualquier caso.

El propósito final, por tanto, será la creación de un protocolo que recoja los estándares mínimos para poder entrenar y utilizar la TA o un motor propio de TA en un entorno de producción y en lenguas con recursos reducidos.

CAPÍTULO II: El multilingüismo y la comunicación digital

En el panorama actual de la comunicación digital, el multilingüismo se presenta como un fenómeno de creciente relevancia que redefine las dinámicas de la interacción en línea. Este capítulo se adentra en el vasto y complejo terreno del multilingüismo y la comunicación digital, con un enfoque particular en los mensajes cortos y su manifestación en la plataforma de Twitter. El capítulo se estructura en dos apartados fundamentales, cada uno de los cuales aborda aspectos clave en la comprensión de esta dinámica comunicativa emergente.

El primer apartado se aproxima al multilingüismo desde tres perspectivas diferentes. En primer lugar, se examina la relación del multilingüismo con las lenguas minorizadas. Luego, se explora la intersección entre las lenguas minorizadas y las nuevas tecnologías, revelando cómo las comunidades que hablan lenguas en riesgo pueden beneficiarse de las plataformas digitales para preservar y revitalizar sus lenguas. Finalmente, se contextualiza el gallego, enfatizando la importancia de la preservación lingüística en un entorno globalizado.

El segundo apartado se sumerge en el análisis de la plataforma de Twitter, por un lado, y en la caracterización del género textual del tuit, la identificación de los componentes principales del tuit y una exploración de los desafíos que plantea la traducción automática neuronal (TAN) en la traducción de mensajes cortos, por otro.

En conjunto, este capítulo busca arrojar luz sobre las interacciones entre el multilingüismo y la comunicación digital en el entorno de los mensajes cortos, ofreciendo una perspectiva detallada y crítica de esta faceta esencial de la comunicación contemporánea.

1. El multilingüismo

El presente apartado versa sobre el multilingüismo y cómo se trata de preservar desde las instituciones públicas. Debido a que una de las lenguas de trabajo de esta investigación es una lengua minoritaria, nos parece muy pertinente contextualizar su estado actual,

especialmente, en lo referente a su presencia en Internet, dado que uno de los propósitos de esta investigación es poder crear un instrumento que sirva de ayuda en la protección y fomento digital. Tal y como invita Torres-Hostench (2023: 7), entendemos desde una perspectiva tecnológica la definición de multilingüismo de la Comisión Europea, esto es, «*the ability of societies, institutions, groups and individuals to engage, on a regular basis, with more than one language in their day-to-day lives*» (Comisión Europea, 2007).

En el marco de esta tesis, son tres las aproximaciones que hemos tenido en cuenta y que articulan los subapartados principales de esta sección: el multilingüismo y las lenguas minoritarias, las lenguas minoritarias y las nuevas tecnologías y, por último, la contextualización del gallego dentro de este marco. En el primer apartado, abordaremos las políticas que han desarrollado las principales instituciones internacionales para proteger y fomentar las lenguas habladas en el mundo. Por su parte, en el segundo apartado, relacionaremos las nuevas tecnologías con las lenguas minoritarias y describiremos los principales retos que plantean para estas lenguas y también las posibilidades que les brindan especialmente en el ámbito de los medios de comunicación. Finalmente, en el último punto, centraremos nuestra atención en la situación actual del gallego teniendo en cuenta lo descrito en los apartados anteriores.

1.1. El multilingüismo y las lenguas minoritarias

Hoy en día, se hablan alrededor de 7000 lenguas en todo el mundo y está ampliamente acordado dentro de la comunidad de expertos que prácticamente el 90 % de esas lenguas serán absorbidas por otras dominantes a finales de siglo. Según el atlas mundial de lenguas de la Unesco (disponible en https://en.wal.unesco.org/), podemos encontrar recogidas cerca de 3000 lenguas en peligro, aunque dicha cifra solo representa las lenguas que el organismo pudo contabilizar en hablantes. Actualmente, se está trabajando en la memoria del mapa para poder tener un reflejo más real del panorama lingüístico. Por esta razón, a lo largo de las últimas décadas se han ido promoviendo distintas iniciativas para proteger la diversidad lingüística.

Tal y como explica el Consejo de Europa (2010) en su publicación *Minority language protection in Europe: into a new decade*, el sistema incipiente de la Sociedad de Naciones ya contaba con ciertas disposiciones relacionadas con la protección de las minorías y de sus lenguas regionales, especialmente enfocadas al derecho a la educación en las respectivas lenguas maternas. A lo largo de los años posteriores distintas declaraciones

universales recogerán artículos relacionados: Declaración Universal de los Derechos Humanos (ONU, 1948) en los artículos 2 y 26; Pacto Internacional de Derechos Civiles y Políticos (ONU, 1966) en su artículo 27; Pacto Internacional de Derechos Económicos, Sociales y Culturales (ONU, 1966) en su artículo 13; o la Convención de la Unesco (1960) contra la Discriminación en la Educación en su artículo 5. Posteriormente, en 1992, las Naciones Unidas adoptaron la Declaración sobre los Derechos de las Personas Pertenecientes a Minorías Nacionales o Étnicas, Religiosas y Lingüísticas. Cuatro años más tarde, en 1996, en la Conferencia Mundial de Derechos Lingüísticos en Barcelona, se elaboró la Declaración Universal de Derechos Lingüísticos, que serviría como modelo para las siguientes convenciones. Finalmente, en los años siguientes, la Unesco fue llevando a cabo distintas iniciativas para el desarrollo de políticas internacionales al respecto, como la Declaración Universal de la Diversidad Cultural, en 2001, que ofrece un marco conceptual para la promoción de la diversidad lingüística y cultural en peligro de extinción. Fruto de esta Declaración, un año más tarde, la Asamblea General de las Naciones Unidas definió en febrero de 2002 el multilingüismo como «medio de promover, proteger y preservar la diversidad de idiomas y culturas en el mundo» y reconoce que el multilingüismo «promueve la unidad en la diversidad y la comprensión internacional» (ONU, 2002: 1).

En octubre de 2003, los Estados Miembros de la Unesco adoptaron las Recomendaciones sobre la Promoción y el Uso del Plurilingüismo y el Acceso Universal al Ciberespacio, de la que hablaremos en los siguientes apartados.

Dentro del marco europeo también se ha estado luchando por proteger el multilingüismo, que se encuentra en la propia definición de Europa. Sin embargo, habrá que esperar a 1981 para encontrar una verdadera apuesta por la protección lingüística, ya que en la época posterior a la II Guerra Mundial no se consideró este tema como una prioridad. Fue en 1981 cuando el Consejo de Europa adoptó la Recomendación 928 sobre los problemas educativos y culturales de las lenguas minoritarias y dialectos en Europa. En este texto no solo se reconocían las identidades lingüísticas como un elemento de desarrollo de Europa, sino que también se recogían medidas que había que implantar, como, por ejemplo, el uso de topónimos originales, el uso de dialectos y lenguas maternas en las escuelas, el apoyo a las lenguas minoritarias y su uso en la educación superior y en medios e instituciones locales. Esta recomendación fomentará posteriormente la Carta Europa sobre Lenguas Regionales y Minoritarias, aprobada en 1992 y con entrada en vigor en 1998. La Carta se centra en la necesidad de proteger el amplio legado lingüístico europeo, al que contribuyen

las distintas lenguas regionales y minoritarias. Este objetivo queda ya reflejado en el preámbulo de esta (recogida en el BOE 222 de 2021, 34734):

> *Los Estados miembros del Consejo de Europa, signatarios de la presente Carta, conscientes del hecho de que la protección y el fomento de las lenguas regionales o minoritarias en los distintos países y regiones de Europa representan una contribución importante a la construcción de una Europa basada en los principios de la democracia y de la diversidad cultural, en el ámbito de la soberanía nacional y de la integridad territorial.*

Además de establecer distintos niveles de protección para las lenguas, la Carta también establecía un sistema de control. Por tanto, cada uno de los países es libre de escoger a medida el conjunto de compromisos que se corresponden con la situación de cada una de las lenguas regionales o minoritarias de su país. Así, un país que deseara proporcionar una protección más práctica podía elegir compromisos específicos siguiendo este sistema de menús. Como resultado, la implantación de la Carta no se realizó de la misma forma en todos los países miembros, ya que algunos ni siquiera la ratificaron.

En cuanto al sistema de control, este partía de un informe nacional inicial que se debía presentar al Secretariado General del Consejo de Europa durante el primer año tras la entrada en vigor de la Carta. Posteriormente, los países deberán presentar un nuevo informe cada tres años. Un comité de expertos designado por el Consejo de Europa examina los informes nacionales y redacta las conclusiones de la evaluación junto con unas propuestas. Posteriormente, el secretario general del Consejo de Europa presenta un informe bienal a la Asamblea Parlamentaria.

El Estado español suscribió la Carta Europea en 1992 y la ratificó el 2 de febrero de 2001, por lo que el primer informe data del año 2002. Nos centraremos en el análisis de los informes en los siguientes apartados.

Además de la Carta, dentro de las instituciones europeas también se han llevado a cabo distintas iniciativas:

- El año europeo de las lenguas (2001)
- Aprobación por parte del Consejo Europeo de una resolución en 2002 para promover la diversidad lingüística y el aprendizaje de lenguas.
- La Comisión Europea lleva financiando desde 1981 diversas acciones de protección y promoción de lenguas minoritarias. Por ejemplo, el plan de acción para promover

el aprendizaje de idiomas y la diversidad lingüística (COM(2003) 449); Mercator (2004), una red europea de centros para la promoción de la diversidad lingüística financiada a través del Programa de Aprendizaje Permanente (2007-2013).

- Publicación en 2005 de una nueva estrategia marco para el multilingüismo, en la que se resalta la importancia de las lenguas minoritarias y regionales en la diversidad lingüística (COM(2005) 596).
- En 2008, la Comisión adoptó la Comunicación sobre multilingüismo (COM(2008) 566).

Junto con las normas y recomendaciones también se han financiado distintas investigaciones sobre el estado y la promoción de las lenguas minoritarias (Pasikowska-Schnass, 2016: 10):

- Carta de París de la ONU en la que se consideran las lenguas como material intangible o bien cultural de la humanidad
- Euromosaic
- Atheme
- ELDIA (European Language Diversity for All)
- Riches (Renewal, Innovation and Change: Heritage and European Society)
- NLPD (Network to Promote Linguistic Diversity)
- RML2future
- La Civil Society Platform on Multilingualism, que busca nuevas ideas y sugerencias en el campo del multilingüismo y que prosperó en el Proyecto de dos años, Poliglotti4.eu.
- Día Europeo de las Lenguas

1.2. Lenguas minoritarias y las nuevas tecnologías

Como ya mencionamos en la introducción del punto anterior, tanto las lenguas regionales y minoritarias como las lenguas menos usadas están en peligro de extinción, y, este peligro se puede ver potenciado por la tecnología digital. Todos los organismos internacionales mencionados tienen a bien señalar que el futuro de muchas de estas lenguas depende de su presencia en los nuevos medios digitales. Si cada vez más población joven se comunica y se informa a través de Internet y si esta información solo está disponible en las lenguas más dominantes, las lenguas menos empleadas o con menos recursos tendrán una muerte digital (Parlamento europeo, 2020). A pesar de esta amenaza, los nuevos medios digitales también

pueden servir como una herramienta para promover la lengua a través de la educación en línea, el aprendizaje de lenguas en línea y las tecnologías de la lengua. La viabilidad de muchas lenguas en el siglo XXI, por tanto, pasa por poder ofrecer un contenido global similar al que se puede consultar en las lenguas más dominantes. Además, este contenido según el informe del Comité de Expertos del Consejo de Europa de 2019, *New technologies, new social media and the European Charter for Regional or Minority Languages*, (2019:16), tiene que ser frecuente y consistente y estar dirigido a los jóvenes. La Red Europea para la Diversidad Lingüística (NPLD) propuso en 2015 la Hoja de Ruta Europea para la Diversidad Lingüística. Esta se articula en cuatro líneas de trabajo: adoptar una política inclusiva del multilingüismo; trabajar para que las lenguas tengan una función de cohesión social y contribuyan al desarrollo económico; usar las tecnologías de la información y la comunicación para el aprendizaje y la promoción de las lenguas y que se dé apoyo a las lenguas regionales o minoritarias.

Posteriormente, el Parlamento Europeo adoptó la resolución sobre igualdad lingüística en la era digital, a la que le siguió un estudio en 2017: *Equality in the digital age. Towards a Human Language Project*, centrado en las barreras idiomáticas dentro de un mercado único europeo lingüísticamente fragmentado y en el que se afirma que el mosaico idiomático europeo supone un desafío para las tecnologías digitales, ya que competidores como Norte América o Asia superan a la UE. Así, se insta a la Comisión a crear un centro de diversidad lingüística que defina unos recursos lingüísticos mínimos para todas las lenguas europeas, entre los que se puede encontrar conjuntos de datos, glosarios, grabaciones de discursos, memorias de traducciones y contenido enciclopédico. Para poder llevar todo esto a cabo también se busca el establecimiento de una estrategia digital que coordine la investigación, el desarrollo y, finalmente, la financiación en muchos casos insuficiente. Como respuesta, la Comisión ha financiado diversos proyectos como la Coordinación de Recursos Lingüísticos Europeos [1](ELRC) a través de la convocatoria Horizon 2020, CRACKER[2] (*Cracking the Language Barrier Federation*) enfocada a la investigación en traducción automática, la META-NET [3](Alianza Multilingüe Tecnológica de Europa), CLARIN [4](Infraestructura Común de Tecnología y Recursos Lingüísticos) o la ELG[5] (*European*

[1] Disponible en: https://www.lr-coordination.eu/es?lang=es. Fecha de consulta: 2 de septiembre de 2023.
[2] Disponible en: http://cracker-project.eu/. Fecha de consulta: 2 de septiembre de 2023.
[3] Disponible en: http://www.meta-net.eu/. Fecha de consulta: 2 de septiembre de 2023.
[4] Disponible en: https://www.clarin.eu/. Fecha de consulta: 2 de septiembre de 2023.
[5] Disponible en: https://live.european-language-grid.eu/. Fecha de consulta: 2 de septiembre de 2023.

Language Grid, Rehm, 2022) enfocada en la democratización de recursos lingüísticos para el mercado europeo.

1.2.1. Principales desafíos de las nuevas tecnologías

Uno de los principales desafíos que se menciona en el informe elaborado por el comité de expertos sobre las nuevas tecnologías, medios sociales y la Carta Europa es el cambio en los patrones de comunicación de la sociedad actual. Muchas de las recomendaciones de la Carta Europea están basadas en los medios de comunicación tradicionales (radio, prensa y televisión) sin tener en cuenta el impacto de los nuevos medios sociales. Por tanto, las nuevas estrategias tendrán que adaptarse a la nueva realidad social (Consejo de Europa, 2019:15):

> *The traditional under-standing of the Charter provisions was dominated by an image of media arrangements that has changed substantially and progressively over time, and subsequently needs reinterpretation. The future of RMLs and their place in society and societal communication infrastructures are less and less decided by the arrangements of traditional electronic mass media, like radio and television, or by print media. In the long run, it will presumably be new media that decide on the fate of regional or minority languages in mass communication.*

Aunque esta realidad cambia dependiendo del país y grupo demográfico, sí podemos extraer una serie de cambios generales. Por un lado, la consulta de noticias a través de Internet gana cada vez más usuarios en comparación con la prensa impresa en la población adulta, mientras que la población más joven opta por consultar las noticias en redes sociales y otras canales de Internet (Consejo de Europa, 2019: 16). Si comparamos el uso de canales de televisión, se observa una tendencia cada vez más mayor especialmente en los sectores más jóvenes de la población de consumir televisión a la carta, lo que ha provocado una mayor oferta de canales, especialmente privados. Como resultado de estos cambios, cada vez hay una menor frontera entre radio, televisión e Internet, ya que se intenta apoyar el medio principal con distintos tipos de contenido audiovisual y buscar así una mayor participación de los usuarios. Los nuevos medios, por tanto, fomentan la interacción con su audiencia. Relacionado con este punto, también encontramos una nueva característica: ya no se orientan los medios al público en general en su sentido más amplio, sino que buscan orientarse a determinados sectores de la población. Por último, el impacto de la tecnología en los nuevos medios. Aquí, podemos hablar de la traducción automática para acceder a otro tipo de información y foros de debate y de la aparición de las tecnologías del lenguaje

(reconocimiento de voz) a la hora de comunicarte con las máquinas y acceder a la información.

1.2.2. Principales recomendaciones

- *Recomendaciones generales*

Dentro de las recomendaciones propuestas por la UNESCO para la promoción del multilingüismo en Internet (2003), podemos distinguir cuatro ejes principales: generación y ampliación de contenido multilingüe, acceso universal a redes y servicios, desarrollo de contenido de dominio público y equilibrio entre los intereses de los titulares de derechos y los intereses generales. Así, dentro del primer grupo, proponen:

- Fomentar el acceso a recursos que reduzcan las barreras lingüísticas y los intercambios humanos en Internet
- Fomentar la creación de contenido digital propio en las lenguas minoritarias
- Promover de políticas estatales y regionales para promover la enseñanza de idiomas en línea
- Promover la investigación en tecnologías de la lengua
- Creación de un observatorio para el seguimiento de las políticas

Por su parte, los siguientes dos grupos son un conjunto de recomendaciones para legislar el acceso universal a Internet y para poner a disposición de los hablantes contenido público.

Resulta especialmente interesante para esta tesis el aspecto relativo al equilibro de intereses, y aunque ya trataremos este tema en profundidad dentro del marco teórico, podemos resumir las propuestas en la siguiente lista:

- Actualización de legislaciones nacionales sobre derechos de autor y su adaptación a Internet
- Animar a que se apliquen ciertas excepciones a dichos derechos cuando proceda
- Prestar atención a la implantación de nuevas tecnologías y sus efectos en el acceso a la información

- *Recomendaciones en el marco legal*

Dado el cambio dramático en el escenario mediático, una de las principales recomendaciones es la adaptación de la Carta Europea al nuevo contexto y,

específicamente, la actualización de los artículos 7, 11, 12 y 14. El informe de 2019 del Consejo Europeo resume así las acciones necesarias en los distintos artículos (2019:50):

> *Article 7*
>
> *Part III undertakings should be interpreted in the light of the overarching objectives formulated in Article 7.1.c (the need for resolute action to promote RMLs in order to safeguard them).*
>
> *According to the objectives laid down in Article 7, states should pursue an adaptation of facilitation and encouragement of the use of RMLs in a manner that takes the digital shift into consideration.*
>
> *Article 11*
>
> *Where media services are migrated to digital platforms, fully or partially, the relevant undertakings of Article 11 cannot be fulfilled if this is done to the detriment of regularity of production, reach and range of distribution, accessibility and visibility of content in RMLs.*
>
> *States should be attentive to the importance of language technologies including digital translations and digital language tools in RMLs, which underpin resources for media production and social media communication in RMLs (also applicable to Article 12).*
>
> *With the spread of new formats of audiovisual media in the digital realm, there is a growing need to support the production of audio and audiovisual works in RMLs (also applicable to Article 12).*
>
> *Training of journalists, staff of media, including freelancers and other producers of media content in RMLs, should develop relevant skills to provide for adequate competences in the digital realm.*
>
> *In the reporting and monitoring processes there should be clarity on the distinction between statuary regulated 'broadcasting'and'online distribution' of audio and audiovisual material and its derivatives in other formats.*
>
> *Article 12*
>
> *States should be made be aware of obstacles in social media contexts, in which RMLs are not supported by the main channels of distribution, and should address this issue in ways that are relevant for this article (also applicable to Article 11).*
>
> *See also the recommendations under Article 11, which are relevant to Article 12.*
>
> *Article 13*
>
> *States should be made be aware of their duty to secure transfrontier availability to services in the digital realm, by addressing geo-blocking and other obstacles that prevent access to media content in RMLs by, for instance by negotiating with copyright holders (also applicable to 11.2.).*

Como podemos apreciar, se busca no sólo adaptar el significado de los conceptos sino también cambiar el enfoque del seguimiento para poder dar una mejor muestra de la realidad en constante cambio. Resulta primordial la recogida y ampliación de datos específicos de las nuevas realidades y que además estén asociados a los grupos demográficos.

1.2.3. Redes sociales y multilingüismo

Como hemos visto anteriormente, una de las mayores revoluciones en el mundo digital es la aparición de la figura consumidor-creador de contenido digital, especialmente ligado a los sectores más jóvenes de la población y a las llamadas redes sociales. El usuario ya no sólo recibe o consume este tipo de información, sino que también crea y participa. A este usuario se le ha venido designado como prosumidor. A pesar de que contamos con pocas publicaciones que recogen datos de uso de lengua en redes sociales, sí podemos hacernos eco del informe Europe's Languages in the Digital Age (Rehm & Uszkoreit, 2012) en el que nos avisan del peligro del uso exclusivo de las lenguas mayoritarias más dominantes. Esta amenaza a las lenguas minoritarias también queda reflejada en el artículo de Fernández, Eguskiza y Lazkano (2020) sobre el uso de internet y la población joven vasca. Las principales conclusiones que extrajeron los autores a partir de un estudio a 2426 jóvenes fueron que en líneas generales la presencia del euskera tanto en el consumo como en la creación de contenidos audiovisuales es escasa. En ambos casos, el castellano supera al euskera, a pesar de contar con los conocimientos y las herramientas necesarias para poder hacerlo en su otra lengua. El estudio también revela otro dato importante: cuanta mayor es la presencia del euskera en Internet, mayor es la creación y consumo de contenido en euskera. De nuevo, volvemos a encontrar la misma recomendación mencionada en las publicaciones de la UNESO y la Unión Europea relativa a la creación de contenido propio en la lengua minoritaria y a la involucración de la población joven:

> *Pese a que la situación de las lenguas indígenas minoritarias de Latinoamérica dista mucho de la del euskera en el País Vasco, la gestión eficaz de Internet y de las nuevas tecnologías por parte de las instituciones educativas de estos países, junto con el uso de las lenguas minoritarias en los nuevos medios por parte de jóvenes hablantes, puede convertirse en una herramienta clave para su recuperación. Por ello, la producción y difusión de contenidos audiovisuales adecuados para la enseñanza/aprendizaje de las lenguas indígenas, la edición de una Wikipedia en estos idiomas o el establecimiento de foros y comunidades virtuales, pueden universalizar su acceso y contribuir a su transmisión intergeneracional. El futuro de esta comunidad virtual depende de su capacidad para atraer a jóvenes indígenas o no indígenas que quieran formar parte del proyecto y que puedan*

> *hacerlo a través de dispositivos digitales como el smartphone, cuyo acceso hoy parece imparable en todo el mundo a pesar de las diferencias socioculturales y económicas. (Fernández, Eguskiza y Lazkano, 2020: 390)*

En otro estudio sobre las lenguas minoritarias europeas y las redes sociales (Ferré-Pavia *et al.*, 2018), se observa que el 75 % de los medios de comunicación tradicionales (periódicos, radios y televisiones) tienen presencia en las redes sociales, especialmente en Twitter y Facebook, seguidas de Youtube.

Language Community		Social Media Presence		Type of Social Media						
	N	Yes	No	Facebook	Twitter	YouTube	Google+	Instagram	Others (Pinterest, Flickr, Vimeo)	Average of Social Media Accounts
Catalan	757	76.1%	23.9%	93.8%	80.2%	66.1%	19.5%	7.5%	6.9%	2.3
Basque	123	76.4%	23.6%	90.4%	86.2%	30.0%	60.0%	10.0%	0.0%	1.8
Welsh	84	47.6%	52.4%	95.0%	57.5%	100.0%	0.0%	0.0%	0.0%	1.7
Galician	60	90.0%	10.0%	96.3%	83.3%	26.1%	56.5%	8.7%	8.7%	3.4
Irish	12	75.0%	25.0%	88.9%	100.0 %	20.0%	20.0%	40.0%	20.0%	2.7
Breton	12	91.7%	8.3%	100.0%	100.0 %	0.0%	100.0%	0.0%	0.0%	2.5
Frisian	11	81.8%	18.2%	100.0%	100.0 %	0.0%	100.0%	0.0%	0.0%	2.7
Sámi	6	83.3%	16.7%	100.0%	60.0%	0.0%	100.0%	0.0%	0.0%	2.8
Corsican	3	100.0 %	0.0%	100.0%	100.0 %	0.0%	50.0%	50.0%	0.0%	3.0
Scottish Gaelic	2	100.0 %	0.0%	100.0%	100.0 %	0.0%	0.0%	0.0%	0.0%	2.0
Total	1,070	75.0%	25.0%	93.8%	80.7%	57.1%	27.7%	8.5%	6.7%	2.3

Source: Authors. *Note*: The total percentages refer to the total quantitative figure (1,070 = 100%).

Ilustración 12. Presencia en redes sociales (Ferré-Pavia *et al.*, 2018: 1077)

Los autores coinciden en resaltar que el estudio llevado a cabo señala la influencia moderadamente positiva de las redes sociales en la vitalidad de las lenguas minoritarias europeas, pero también recalcan la necesidad de adaptarse constantemente a los cambios tecnológicos, de mejorar la interacción con los usuarios y de generar y actualizar contenido. Esta misma idea también la encontramos en Cunliffe (2018). El autor, además, también señala que el futuro dependerá del interesé de las nuevas generaciones:

> *It seems inevitable that most, if not all minority languages will have to face the opportunities and challenges posed by social media. Social media does appear to offer minority language communities the opportunity to create virtual breathing spaces, but those communities also need the capacity and the desire to do so. The way in which minority language communities respond to these opportunities and challenges will ultimately determine the effect that social media has on the maintenance of their languages. (Cunliffe, 2018:476)*

Un ejemplo de que se puede revitalizar una lengua a través de las nuevas tecnologías es el caso del aragonés. Tal y como explican Martínez-Cortés *et al.* (2010), esta lengua está considerada en peligro de extinción, pero, aun así, la comunidad de hablantes se ha esforzado en los últimos años por crear una de red de información, sitios y recursos que ha dado vitalidad a la lengua y con la que esperan cambiar su curso. Concluyen que es necesario tener en cuenta tres puntos clave:

> *Thus, the three key points: the creation of linguistic resources and software localization, normalization of use of the language on internet, and outreach and education, need to be taken into account by the competent authorities, or be analyzed by associations to address coordinated actions. Martínez-Cortés et al. (2010:8)*

1.3. Contextualización del gallego

En la sección anterior, hemos visto que el principal desafío de las lenguas menos usadas es contar con una presencia continuada en el mundo digital, especialmente, dentro del rango más joven de la población para garantizar su supervivencia. Así, todas las posibles estrategias pasan no sólo por digitalizar y democratizar recursos en Internet sino también por proporcionar herramientas TIC que favorezcan la comunicación en línea de estas herramientas. En esta laboriosa tarea, principalmente impulsada por las instituciones públicas, puede ser de gran ayuda la traducción (González, 2002; González y Díaz, 2007) y, en especial, la traducción automática, ya que no sólo aumentaría el contenido existente, sino que el desarrollo de su actividad favorecería la aparición de toda una serie de herramientas tecnológicas de utilidad para los usuarios finales como diccionarios, terminologías, correctores ortográficos, etc.

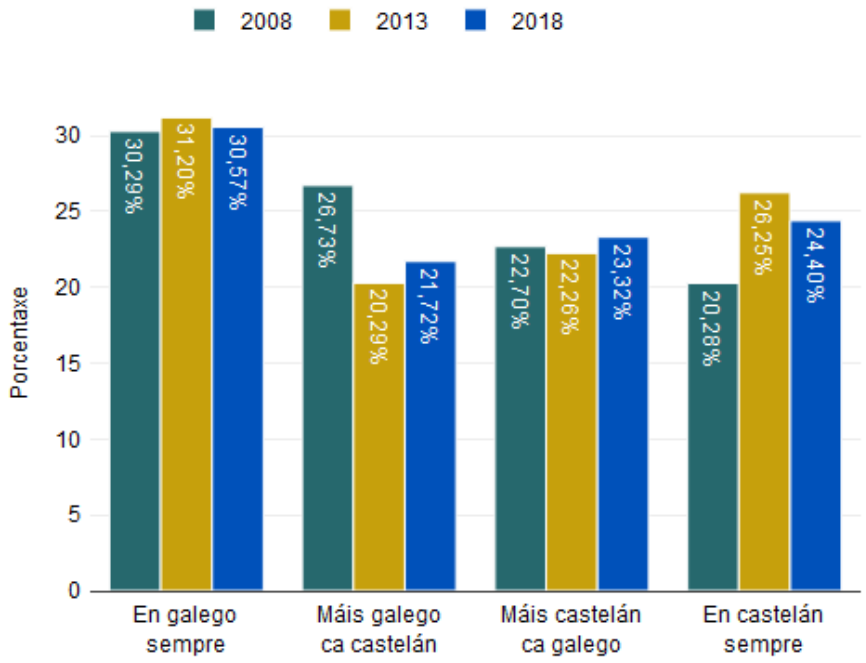
Tal y como indica Henrique Monteagudo (2019: 17-18), dentro del conjunto de lenguas minoritarias del estado español, el gallego presenta unos niveles de compresión y de uso elevados dentro de su población, ya que es la lengua inicial y habitual de casi la mitad de la población gallega y la segunda de un cuarto más de habitantes de Galicia. Sin embargo, a pesar de su oficialización como lengua propia de la Comunidad Autonómica, de su

enseñanza en las escuelas primarias y secundarias, de las políticas de normalización lingüística, de la emisión íntegra en gallego de un canal de televisión autonómico y de una producción cultural notablemente amplia y diversificada, está lejos de salir de una situación de diglosia y está perdiendo cada vez más hablantes jóvenes al constatarse una alarmante brecha intergeneracional.

En los siguientes apartados, recogeremos los datos actuales publicados sobre el gallego, la situación actual recogida en los informes presentados al Consejo de Europa dentro del marco que proporciona la Carta Europea y justificación de cómo promover el gallego dentro del multilingüismo europeo.

1.3.1. Datos demográficos, hablantes y presencia en Internet

Galicia cuenta con 2 701 819 habitantes según datos del *Instituto Galego de Estatística* en 2020. En el informe de 2019 sobre el conocimiento del gallego, podemos apreciar un descenso en el número de hablantes mayoritariamente en gallego, especialmente llamativo en los sectores más jóvenes de la población (véase la Ilustración 2).



[6]

Ilustración 13. Encuesta estructural a hogares (IGE, 2019:5)

Si desglosamos por edad, la tendencia a la baja en las generaciones más jóvenes está claramente diferenciada. Si nos fijamos en la columna *en castelán sempre*, vemos que los

[6] Todos los gráficos directamente extraídos de las fuentes en gallego no se van a traducir.

valores son muy superiores en las generaciones más jóvenes con respecto a la columna que recoge las personas que siempre usan el gallego como lengua habitual.

Persoas segundo a lingua na que falan habitualmente por idade. Galicia. Ano 2018

Unidade: porcentaxe (%)

Idade	En galego sempre	Máis galego ca castelán	Máis castelán ca galego	En castelán sempre
Total	30,57	21,72	23,32	24,40
De 5 a 14 anos	14,27	11,85	29,75	44,13
De 15 a 29 anos	18,94	18,45	30,75	31,86
De 30 a 49 anos	23,66	20,64	28,25	27,45
De 50 a 64 anos	32,78	25,35	21,51	20,36
De 65 ou máis anos	48,48	24,91	12,96	13,65

Ilustración 14. Ilustración del porcentaje de hablantes por edad (IGE: 2019: 4)

Si atendemos ahora a la presencia del gallego en Internet, podemos resaltar que la UNESCO sitúa al gallego dentro de la categoría de lenguas locales de países desarrollados. Con respecto a Internet, la UNESCO no es capaz de determinar el futuro de la lengua, ya que esta se ve amenazada por el inglés y por la lengua nacional dominante respectiva (Pimienta, D., Prado, D. y Blanco, A., 2009: 14).

La presencia del gallego en Internet ha ido descendiendo desde 2003. Si consultamos los datos más recientes (véase Ilustración 15), podemos ver que se usa en menos del 0,1 % del total de páginas web del mundo, lo que lo sitúa en el puesto 61 de las lenguas más usadas en Internet, muy por debajo del euskera y del catalán.

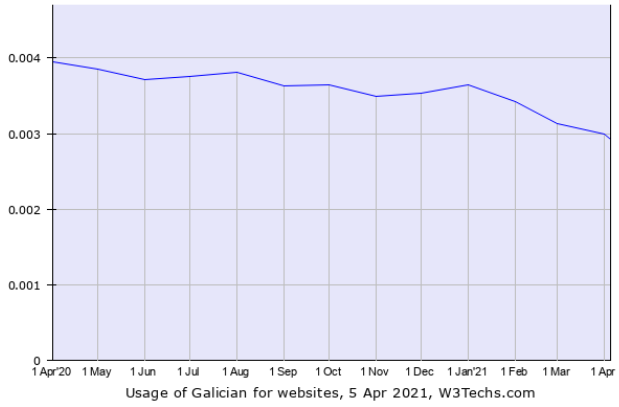


Ilustración 15. Uso del gallego en Internet a 5 de abril de 2021 según W3Techs.com

Si nos fijamos en el número de dominios .gal, vemos que hay registrados 5954 sitios según el portal NTLDstats.com, lo que representa el 0,02 % del total de dominios inscritos en el mundo.

Con respecto a la localización de software, Fernández (2016) afirma que, aunque con retraso respecto a otras lenguas, la mayoría de los usuarios puede disponer de los softwares más empleados en gallego, a excepción de Apple. Aun así, esta oferta viene promovida mayoritariamente por voluntarios (Proxecto Trasno[7], para la traducción de software libre al gallego) y las instituciones. Losada Romero (2016: 13) apunta que una de las asignaturas pendientes son las redes sociales. Domínguez y Ramallo ya observaron en 2012 que el uso mayoritario de los jóvenes en las redes sociales de aquel entonces es el castellano y que hay más demanda de localización de la interfaz de las redes sociales que voluntad de emplear el idioma en las mismas. También existe un gran desconocimiento de redes sociales propias en gallego (Caxigo[8], Chuza[9], etc.).

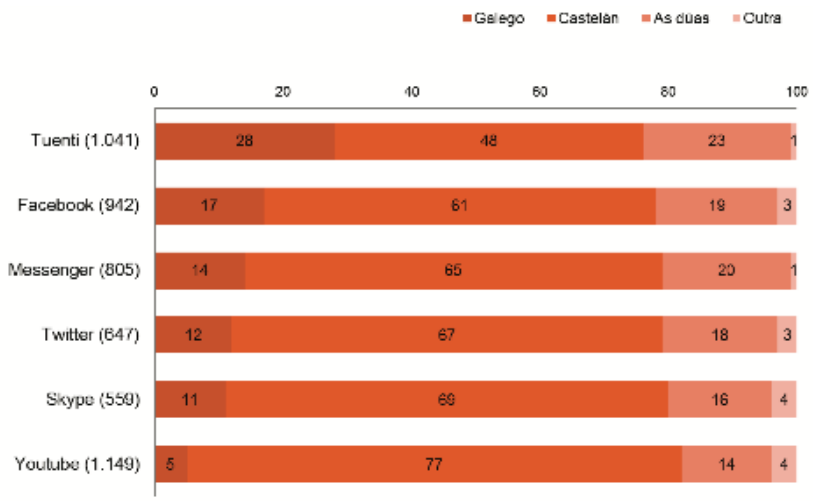


Ilustración 16. Lengua empleada en las redes sociales por la juventud gallega (Domínguez y Ramallo, 2012: 28).

Podemos por tanto afirmar que no se ha cambiado la tendencia observada por el Observatorio da Lingua Galega en un informe de 2009 sobre la situación de la lengua gallega en la sociedad en el que se concluye que el idioma mayoritario empleado para consultar información en Internet es el castellano (2009: 75). Tendencia que Casares *et al.*

[7] Accesible en: https://trasno.gal/. Fecha de consulta: 15 de agosto de 2023.
[8] Accesible en: https://caxigo.gal/. Fecha de consulta: 15 de agosto de 2023
[9] Accesible en: http://www.chuza.gal. Fecha de consulta: 15 de agosto de 2023.

vuelven a constatar en su informe sobre la juventud gallega y las pantallas (2021: 44) a partir de los datos del *IGE* y de la *Enquisa estructural a fogares* (2018).

Por último, nos gustaría señalar el análisis que se hizo del gallego en la era digital dentro del proyecto META-NET por García Mateo *et al.* (2012) en el que se presentan las siguientes gráficas y que coloca al gallego siempre en un soporte fragmentario o inexistente:

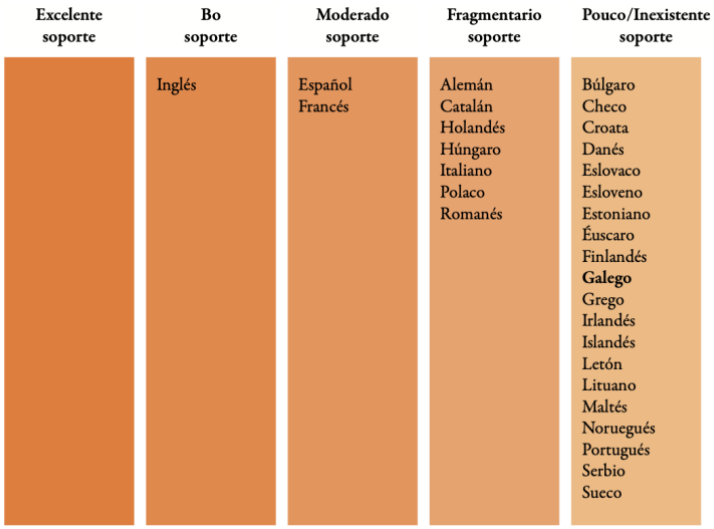


Ilustración 17. Soporte relativo a la traducción automática (García Mateo *et al.*, 2012: 33)

Veamos ahora la gráfica sobre recursos disponibles al alcance de los hablantes en lengua gallega:

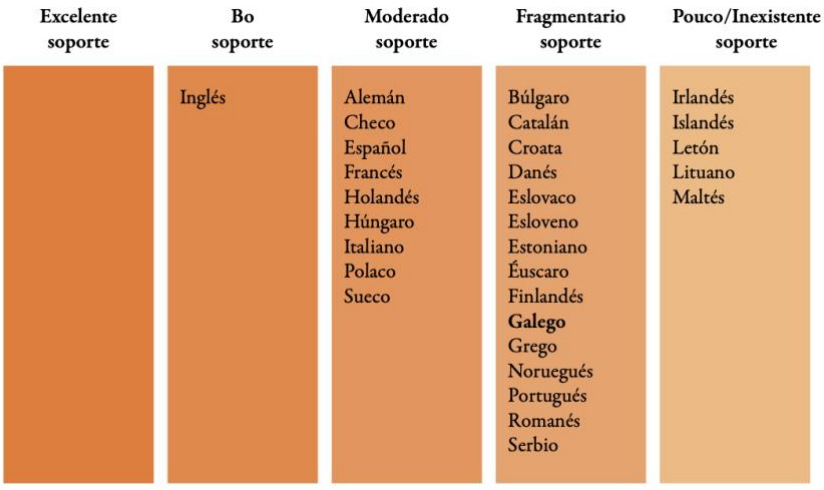


Ilustración 18. Soporte relativo a los recursos disponibles para las tecnologías del lenguaje (García Mateo *et al.*, 2012: 34)

Como podemos apreciar apenas se documentan algunos recursos disponibles en tecnologías del lenguaje, pero no específicamente de traducción automática, lo que respalda nuestra motivación para crear un motor de traducción automática hacia el gallego.

1.3.2. Informes de la Carta Europea sobre el gallego

Como ya se ha mencionado anteriormente, tras la entrada en vigor en España de la Carta Europea de las Lenguas Regionales o Minoritarias en 2001, España debía presentar el primer informe en 2002, el segundo informe 2002-2005 en 2006, y así sucesivamente cada tres años. El informe más actual hasta la fecha de redacción es de 2013-2016 al que le siguen las recomendaciones como respuesta al quinto informe en 2019 y un informe intermedio relativo a las acciones inmediatas en 2020, junto con su ampliación en 2021.

En dichos informes, se realiza un análisis del cumplimiento de las recomendaciones recogidas en la Carta según cada una de las lenguas minoritarias del estado. Debido a que los informes cubren un amplio espectro de temas, nos vamos a centrar en lo establecido en los informes más actuales sobre el gallego dentro del área del multilingüismo y las nuevas redes sociales, ya que es el tema planteado en esta tesis.

Dado que el informe sigue el modelo de evaluación planteado por la Carta, nos encontramos información específica relacionada con los nuevos medios sociales (aspecto ya mencionado anteriormente, al hablar de los desafíos de las nuevas tecnologías). El apartado correspondiente a los medios de comunicación se centra exclusivamente en el concepto tradicional de los medios: radio, televisión y prensa. Sí se menciona la traducción automática al gallego de periódicos en su versión digital como La Voz de Galicia, el Correo Gallego o El Progreso. Aun así, no hay mención a la presencia o difusión del gallego en las nuevas redes sociales.

Otro dato relevante del documento es la mención a las actividades relacionadas con la traducción y la investigación terminológica. Aquí se menciona la labor de la Asesoría Lingüística adscrita a la *Secretaría Xeral de Política Lingüística* (SXPL) en lo relativo al apoyo en materia de asesoramiento lingüístico y producción y oferta de recursos tecnologías para el uso de la lengua gallega; y del *Servizo de Terminoloxía en Galego* (Termigal). También se menciona que la SXPL forma parte, como miembro fundador y de pleno derecho, de la *Network To Promote Linguistic Diversity* (NPLD[10]) y participa directamente en las iniciativas de esta red asociativa de ámbito europeo para la promoción de la diversidad lingüística.

[10] Accesible en: http://www.npld.eu/. Fecha de consulta: 15 de agosto de 2023

En la ampliación del informe intermedio de 2021, se observa una falta de traducciones de documentos oficiales y un acceso limitado a recursos para garantizar los derechos de los ciudadanos en lengua gallega, tal y como muestra la siguiente gráfica.

GALICIA	Highly insufficient	Insufficient	Improvable	Acceptable	Satisfactory
Translation of documents available to citizens					
The guarantee that citizens will be answered in the co-official language of their choice in their dealings with the Government Delegation					
The resources available to guarantee citizens' language rights are considered sufficient by the Government Delegation					

Ilustración 19. Acceso a recursos por parte de los ciudadanos en lengua gallega

Por ejemplo, aunque se ha hecho una gran labor traduciendo las principales páginas web institucionales, el informe pone de manifiesto que en la mayoría de los casos sólo se han traducido la página de inicio (página web de la Agencia Tributaria) o que las traducciones no llegan a todo el mapa del sitio (página web de la Seguridad Social).

1.4. Consideraciones finales

Tras lo expuesto en los apartados anteriores, queda demostrado que la irrupción de las nuevas tecnologías supone al mismo tiempo una ventaja y una amenaza para las lenguas minoritarias. En el caso particular que nos ocupa, hemos podido observar la necesidad de una mayor presencia del gallego en los nuevos medios de comunicación, tanto en los contenidos propios y la tecnología como en la interacción de los usuarios. Por esta razón, se ha querido desarrollar la investigación dentro de las nuevas redes sociales, no sólo por la falta evidente de bibliografía sino también siguiendo el espíritu de los principales organismos internacionales de promoción del multilingüismo.

2. La comunicación digital en mensajes cortos

En este apartado, nos centraremos, por un lado, en el género textual, esto es, en la caracterización de la comunicación digital en mensajes cortos y, por otro lado, en la plataforma de Twitter, hoy en día X, dado que es una de las redes sociales más populares y se ha convertido en una herramienta clave para la comunicación instantánea y la difusión de información en todo el mundo.

Los mensajes en Twitter, también conocidos como tuits, inicialmente se limitaban a un máximo de 280 caracteres, lo que los hace ideales para la comunicación rápida y efectiva. Si bien hoy en día ya no existe esta restricción, sigue siendo lo habitual entre los usuarios. Debido a esta limitación en la longitud, los usuarios de Twitter han desarrollado un estilo de escritura único y conciso, que se ha vuelto cada vez más popular y ha llevado a la creación de una nueva forma de lenguaje y comunicación.

En este sentido, es importante caracterizar la comunicación digital en mensajes cortos en Twitter, para entender cómo se está desarrollando esta nueva forma de comunicación y cómo se está utilizando el lenguaje en esta plataforma. Además, la comprensión de estas características es esencial para la evaluación efectiva de los sistemas de traducción automática para Twitter, ya que la comprensión de las particularidades del lenguaje y la comunicación en esta plataforma es esencial para el desarrollo de modelos de evaluación de la traducción precisos y eficaces.

Por lo tanto, en este apartado, se presentarán las principales características de la comunicación digital en mensajes cortos en Twitter, incluyendo su estilo de escritura, la forma en que se emplean los *hashtags*, los emoticonos y las menciones, y la importancia del contexto en la comprensión de los tuits. Además, se analizarán los desafíos específicos que plantea la traducción automática en Twitter y se explorarán las principales técnicas usadas para abordar estos desafíos. En resumen, este apartado proporcionará una comprensión detallada de las características de la comunicación digital en mensajes cortos en Twitter y su relevancia para la traducción automática.

2.1. Twitter a lo largo de los años

La historia de Twitter se remonta a marzo de 2006, cuando Jack Dorsey, Noah Glass, Biz Stone y Evan Williams fundaron la empresa Odeo, una plataforma de pódcast. Sin embargo, tras la disminución de la popularidad de los pódcast, se vieron obligados a replantearse su enfoque. Fue en este contexto que Jack Dorsey propuso la idea de una plataforma de microblogueo, lo que dio lugar al nacimiento de Twitter. Desde su lanzamiento público en julio de 2006, Twitter experimentó un crecimiento constante, convirtiéndose en un lugar donde las personas podían compartir sus pensamientos, noticias y eventos en tiempo real. Durante esta etapa, la plataforma se centró en establecer las características básicas que la definirían, como el concepto de tuits, seguidores y retuits. En los años sucesivos, Twitter añadirá más funcionalidades a la plataforma como imágenes, GIF, vídeos o *hashtags*.

En octubre de 2022, Elon Musk compra la plataforma e inicia una serie de cambios profundos en la plataforma como, por ejemplo, la posibilidad de escribir hasta 4000 caracteres para usuarios de Twitter Blue o el cambio de marca a *X*. A los efectos de esta tesis, seguiremos refiriéndonos a la plataforma como Twitter, dado que la investigación se llevó a cabo antes del cambio de marca y de cualquiera de sus funcionalidades futuras. Asimismo, nos centraremos en el tuit clásico porque sigue siendo el uso más habitual.

2.2. El género textual del tuit

El tuit es una forma de comunicación digital en mensajes cortos que se ha popularizado en los últimos años y ha creado un nuevo género textual. Según Leppänen y Kytölä (2016), el tuit es un género híbrido que combina características de la escritura literaria y la comunicación oral, y que ha desarrollado sus propias convenciones lingüísticas y discursivas.

En términos de género, el tuit se caracteriza por su brevedad, su informalidad y su capacidad para transmitir información de manera rápida y efectiva. En este sentido, el tuit se ha convertido en un género fundamental para la comunicación en línea y ha transformado la forma en que las personas se comunican e interactúan en la era digital.

Además, el tuit también se caracteriza por su capacidad para generar conexiones sociales y promover la participación ciudadana. Como señala Papacharissi (2010), el tuit es una forma de microblogueo que permite a los usuarios publicar actualizaciones sobre sus vidas,

compartir noticias y opiniones, y conectarse con otros usuarios a través de la plataforma. De esta manera, el tuit se ha convertido en una herramienta importante para la creación de comunidades en línea y para la participación ciudadana en la política y la cultura (González-Bailón *et al.*, 2012).

2.3. Componentes de un tuit

Un tuit es un mensaje corto que se compone de varios elementos clave que lo definen.

- Texto: el texto es el contenido del mensaje y es limitado a un máximo de 280 caracteres en la actualidad. El texto puede incluir *hashtags*, menciones, enlaces y emojis para enriquecer el mensaje.
- *Hashtags*: los *hashtags* son etiquetas que se utilizan para categorizar el contenido y hacerlo más fácilmente accesible a otros usuarios en la plataforma. Los *hashtags* se escriben con el símbolo # seguido de una palabra o frase sin espacios.
- Menciones: las menciones son una forma de dirigirse directamente a otro usuario en la plataforma. Las menciones se escriben con el símbolo @ seguido del nombre de usuario de la persona a la que se quiere mencionar.
- Enlaces: los enlaces son URL que se usan para enlazar el contenido externo al tuit. Los enlaces pueden ser acortados para que ocupen menos espacio en el tuit.
- Emojis: los emojis son imágenes pequeñas que se emplean para expresar emociones o ideas de manera visual. Los emojis se escriben utilizando caracteres especiales y se pueden utilizar para enriquecer el contenido del tuit.
- Adjuntos: contenido multimedia o texto

Según Danet (2020), el tuit es un género discursivo que se compone de varios elementos clave que lo definen. El texto es el elemento central del tuit y su limitación de caracteres es lo que hace que este género sea único. Los *hashtags* y las menciones se han convertido en una forma efectiva de categorizar el contenido y establecer conexiones sociales en la plataforma. Los enlaces son una forma de compartir información externa y los emojis son una forma de expresar emociones y añadir contexto al mensaje.

2.4. Principales desafíos para traducir automáticamente tuits

El principal problema de la traducción automática en el contexto de los tuits es la naturaleza del género discursivo. Como ya hemos mencionado anteriormente, los tuits son mensajes cortos y concisos, que a menudo se escriben empleando lenguaje coloquial, jerga y

abreviaturas. Además, los tuits a menudo incluyen *hashtags*, menciones, enlaces y emojis, lo que añade una capa adicional de complejidad a la traducción automática.

Zappavigna (2012) concluye que la brevedad y la densidad del lenguaje en los tuits hacen que la traducción automática sea un desafío. Asimismo, los tuits a menudo incluyen neologismos y términos de moda que pueden ser difíciles de traducir con precisión. Por lo tanto, las traducciones automáticas de los tuits pueden producir resultados inexactos y malinterpretaciones del mensaje original.

Otro problema que se presenta en la traducción automática de los tuits es la presencia de jerga y abreviaturas. Según Choudhury *et al.* (2010), los usuarios de Twitter a menudo usan jerga y abreviaturas para ahorrar espacio y tiempo en la escritura de los mensajes. Sin embargo, estas abreviaturas pueden ser confusas para los sistemas de traducción automática y pueden dar lugar a traducciones inexactas.

Además, los *hashtags* y las menciones son elementos clave en los tuits que permiten a los usuarios categorizar el contenido y establecer conexiones sociales en la plataforma. La traducción automática de los *hashtags* y las menciones puede ser especialmente difícil debido a que no siempre tienen una traducción directa en otros idiomas. Según Lommel (2013), la traducción automática de los *hashtags* y las menciones puede ser imprecisa y puede ocasionar una pérdida de contexto para los usuarios.

Por último, Lohar *et al.* (2019) definen dos problemas adicionales: ruido lingüístico, es decir, incorrecciones en cuanto a la norma lingüística de la lengua que no suponen problemas de entendimiento para los humanos, pero sí para los motores de traducción automática y la escasez de contenido bilingüe, necesario para el entrenamiento de los motores.

2.5. Consideraciones finales

En conclusión, la traducción automática de los tuits es un desafío debido a las particularidades del género discursivo tanto lingüísticas como paralingüísticas (emoticonos, *hashtags*, etc.). Los tuits son mensajes cortos y densos, que a menudo utilizan lenguaje coloquial, jerga y abreviaturas. Además, los *hashtags* y las menciones son elementos clave en los tuits que permiten a los usuarios categorizar el contenido y establecer conexiones sociales en la plataforma.

Estas particularidades hacen que la traducción automática de los tuits requiera una estrategia que tenga en cuenta estas características específicas. La estrategia pasa por incluir el uso de herramientas de preprocesamiento de los datos para identificar y manejar los elementos específicos de los tuits, como las menciones, *hashtags* y emojis.

Además, la estrategia de entrenamiento de un motor de traducción automática para redes sociales debe incluir el empleo de modelos de lenguaje específicos del género discursivo. Estos modelos pueden ser entrenados con un corpus de tuits para mejorar la precisión de la traducción automática de este tipo de contenido, tal y como veremos en las siguientes secciones.

En resumen, la traducción automática de los tuits requiere una estrategia específica para manejar las particularidades del género discursivo. La identificación y el manejo adecuado de los elementos específicos de los tuits, así como el uso de modelos de lenguaje específicos del género, pueden mejorar significativamente la calidad de la traducción automática de los tuits, tal y como se podrá ver en los siguientes capítulos.

CAPÍTULO III: La traducción automática neuronal

La traducción automática (TA) es una de las aplicaciones más relevante del procesamiento del lenguaje natural (PLN). La creciente globalización y el aumento de la comunicación internacional han hecho que la traducción automática sea una necesidad en muchas situaciones, tanto personales como profesionales.

La TA se refiere al proceso de traducir automáticamente el contenido de un idioma a otro utilizando sistemas de inteligencia artificial. Esta tecnología ha evolucionado significativamente en los últimos años gracias al desarrollo de algoritmos de aprendizaje automático y la disponibilidad de grandes conjuntos de datos para entrenamiento. Gracias a estos avances, surgió la traducción automática neuronal (TAN, por sus siglas en español) que, como veremos en los siguientes apartados, recibe su nombre por emplear redes neuronales profundas para realizar el proceso de traducción.

En este capítulo, se abordará directamente la TAN, ya que es la tecnología que se usó en esta investigación. Primero se explicará cómo funciona, posteriormente cuáles son las arquitecturas principales y finalmente cuáles son los procedimientos de entrenamiento más comunes.

1. Funcionamiento de la traducción automática neuronal

La TAN es un enfoque revolucionario en la traducción automática que ha logrado resultados impresionantes en términos de precisión y fluidez en la traducción de textos. A diferencia de los modelos de traducción automática estadísticos anteriores, que se basaban en reglas gramaticales y en la comparación de las probabilidades de ocurrencia de palabras en diferentes idiomas, la TAN usa redes neuronales profundas para traducir textos de un idioma a otro. Los sistemas de TAN se basan en redes neuronales artificiales que se componen de miles de unidades artificiales extraídas del corpus empleado para su

entrenamiento (Forcada, 2017). Tal y como recogen Casacuberta *et al.* (2017: 69), una red neuronal se compone de un conjunto de unidades de procesamiento simple (o neuronas artificiales) densamente conectadas entre sí y cuya función es realizar un producto escalar entre las entradas a la neurona y un vector de pesos (asociado a cada neurona) seguido de una función no lineal de activación. Además, en la TAN las unidades de palabras o subpalabras se procesan de forma paralela y distribuida para formar representaciones de palabras y su contexto. Forcada (2017) define una representación como una representación de la activación de los estados de cada red neuronal en un grupo específico, llamado capa. Estas representaciones son las que producen la traducción. Por tanto, la TAN funciona primero formando representaciones distribuidas del corpus de entrenamiento, prediciendo la siguiente palabra y aprendiendo de las representaciones de palabras para formar representaciones de unidades más grandes. En la práctica, esto significa que la TAN primero codifica y luego decodifica.

La arquitectura de la TAN consta de dos componentes principales: el codificador y el decodificador. El codificador procesa la secuencia de entrada (texto en el idioma origen) y la representa en un espacio de alta dimensionalidad, es decir, transforma la frase original en una lista de vectores a los que asigna un símbolo. El decodificador, a su vez, utiliza la representación generada por el codificador para generar la secuencia de salida (texto en el idioma destino).

Tal y como explica do Campo *et al*. (2019:7), para la asignación de valores, se usan letras minúsculas en negrita para identificar los vectores, letras mayúsculas en negrita para las matrices, letras mayúsculas en cursiva para representar conjuntos, letras mayúsculas para secuencias y letras minúsculas para presentar los símbolos individuales de cada secuencia. Si, por ejemplo, X e Y representan una oración de origen y de llegada, $X = \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_M$ sería la secuencia de símbolos M presentes en la oración original e $Y = \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N$, la secuencia de símbolos N de la oración de llegada. Por tanto, el codificador realizaría la siguiente función: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ = codificadorRNN($\mathbf{x}_1, \ldots, \mathbf{x}_M$). En esta ecuación, $\mathbf{x}_1, \mathbf{x}_2$, etc. son la lista de vectores. El número de unidades de la lista se corresponde con el número de símbolos de la oración original. Así, utilizando la regla de la cadena, la posibilidad condicional de una secuencia P(Y|X) se podría formular así:

$$P(Y|X) = P(Y|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_M)$$
$$= \prod_{i=1}^{N} P(y_i|y_0, y_1, y_2, ..., y_{i-1}; \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_M)$$

Ilustración 20. Fórmula de probabilidad condicional de una secuencia (Wu *et al.*, 2016:3)

Donde $\mathbf{y}_0$ es el símbolo especial de inicio de oración que se coloca en cada frase traducida.

$$P(y_i|y_0, y_1, y_2, y_3, ..., y_{i-1}; \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_M)$$

Ilustración 21. Fórmula de probabilidad de Google Neural Machine Translation (Wu *et al.*, 2016:3)

Durante la inferencia, la probabilidad del siguiente símbolo se calcula a partir de la codificación de la oración original y de la secuencia traducida decodificada obtenidas hasta el momento, tal y como vemos en la Ilustración 21.

2. Tipos de arquitecturas de sistemas TAN

En esta sección describimos las dos principales arquitecturas de sistemas TAN que, según Pérez-Ortiz *et al*. (2022), son dos: las llamadas *transformer* y redes neuronales recurrentes.

2.1. Arquitectura basada en *transformer*

Un sistema *transformer* de TAN (Vaswani *et al.* 2017) se compone de un módulo que procesa incrustaciones contextuales de palabras para cada una de las palabras de la oración de origen y otro módulo que predice sucesivamente cada palabra de la oración de llegada. Hablamos del codificador y decodificador descritos en el punto anterior. Para predecir las palabras en la lengua de llegada, el decodificador presta atención a las incrustaciones de todas las palabras de la frase de origen, así como a las incrustaciones de las palabras de llegada ya generadas. Por tanto, el módulo de atención actúa como nexo de conexión entre los dos módulos y permite al decodificador centrarse en las distintas partes de la oración original mientras está ejecutándose. La atención se logra mediante la asignación de pesos a las diferentes partes del texto de origen, de manera que el modelo se concentra en las partes más importantes durante la traducción.

Pérez-Ortiz *et al.* (2022:158) muestran el ejemplo de traducción automática de una frase usando una arquitectura completa basada en *transformer*.

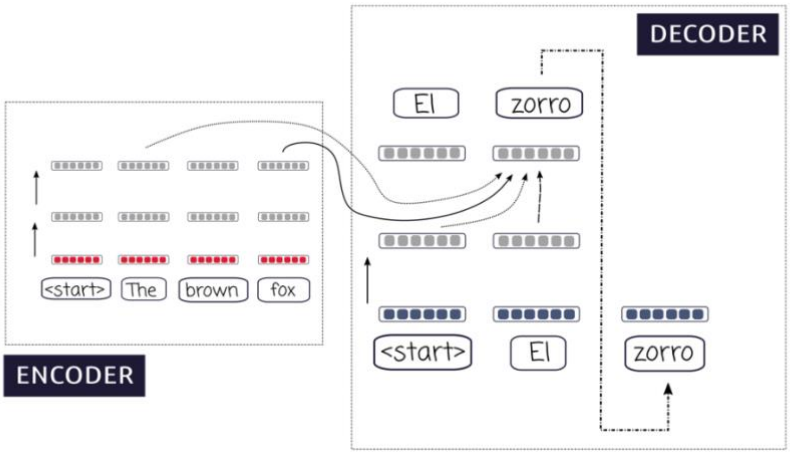


Ilustración 22. Ejemplo de arquitectura completa basada en *transformer* recogido en Pérez-Ortiz *et al.* (2022:158)

Como podemos apreciar, en este ejemplo, se muestra un módulo *encoder* (codificador) con tres capas y se señala la atención necesaria para predecir en el módulo *decoder* (decodificador*)* la traducción «zorro».

A continuación, se puede ver otro ejemplo de arquitectura, esta vez de GNMT, recogida en Wu *et al.* (2016: 4).

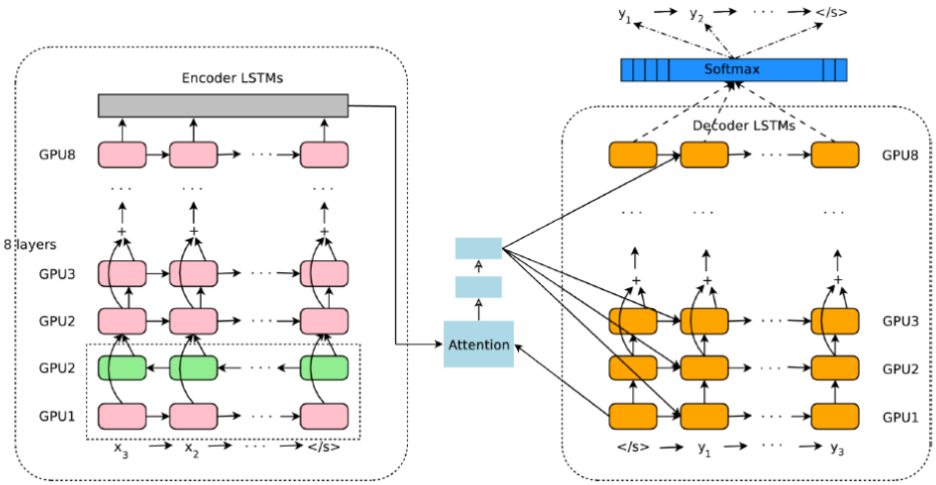


Ilustración 23. Modelo de arquitectura de Google Neural Machine Translation

Como podemos ver en la ilustración anterior y tal y como explican los autores, a la izquierda se encuentra la red del codificador, a la derecha se encuentra la red del decodificador y en medio está el módulo de atención. Dentro de la capa inferior del codificador encontramos nodos bidireccionales: los nodos rosados recopilan información de izquierda a derecha, mientras que los nodos verdes recopilan información de derecha a izquierda. Las otras capas del codificador son unidireccionales. Las conexiones residuales comienzan a partir de la tercera capa desde abajo en el codificador y el decodificador. El modelo se divide en múltiples unidades de procesamiento gráfico (GPU) para acelerar el entrenamiento. Durante el entrenamiento, las capas inferiores del codificador bidireccional calculan en paralelo primero. Una vez que ambas terminan, las capas unidireccionales del codificador pueden comenzar a calcular, cada una en una GPU separada. La capa *softmax* también se divide y se coloca en múltiples GPU.

2.2. Arquitectura basada en redes neuronales recurrentes

Aunque los principales proveedores comerciales de traducción automática están basados en la arquitectura *transformer*, existen otras alternativas. Una de ellas es la arquitectura basada en redes neuronales recurrentes (Bahdanau *et al.* 2015). Esta arquitectura también se compone de un codificador que procesa la frase de origen en incrustaciones de palabras y de un decodificador que predice las palabras de llegada prestando atención. La diferencia está en que tanto el codificador como el decodificador procesan de manera local las incrustaciones de palabras. Por ejemplo, las incrustaciones de la quinta palabra se basan tanto en las incrustaciones de la cuarta palabra como en las de la siguiente palabra, ya que trata la oración de izquierda a derecha y de derecha a izquierda.

3. Proceso de *fine-tuning*

El *fine-tuning* o ajuste fino consiste en la reutilización de un modelo ya entrenado para una tarea específica y el reentreno de ese mismo modelo cambiando ciertos parámetros para la realización de otra tarea. Esta técnica se denomina fina porque se modifican exclusivamente algunas capas o representaciones dándole más importancia para obtener la salida deseada. Miceli *et al.* (2017: 1489) definen esta técnica de la siguiente manera: «*to continue training an existing model which was trained on out-of-domain data within domain training data*». Luong y Manning, 2015; Sennrich *et al.*, 2016 demostraron que esta técnica es efectiva para traducción automática y, como veremos más adelante,

resultará decisiva para nuestro entrenamiento del motor de TA en el contexto de una lengua con pocos recursos.

El *fine-tuning* se puede hacer bien con datos masivos (*over-fiting*) para entrenar el motor existente en una tarea similar usando como base el modelo entrenado y los numerosos nuevos datos para seguir con el entrenamiento; o bien con pocos datos. Para realizar esta técnica con pocos datos, primero se congelan las capas de conocimiento preexistentes y se reentrena las últimas capas con los nuevos datos.

4. Consideraciones finales

En este apartado, nos centramos en describir cómo funciona la TAN, de qué elementos se componen su arquitectura, qué variantes existen y una de las estrategias más empleadas para superar la escasez de datos en un dominio específico. Si bien se han repasado las funciones principales, este tipo de tecnología está en constante evolución.

Actualmente, la TAN es el paradigma actual, a pesar de la irrupción en los últimos años de los modelos grandes de lenguaje (LLM por sus siglas en inglés). Si bien las LLM se impondrán en un futuro especialmente en las lenguas con más recursos, hoy en día, la TAN todavía supone un desafío para las lenguas con pocos recursos. Por esta razón, se eligió esta tecnología para las lenguas de trabajo y se le dedica este apartado en el marco teórico.

CAPÍTULO IV: Corpus y su legalidad

En el contexto de la investigación lingüística y el análisis de lenguaje natural, la creación de corpus representa un elemento fundamental en el desarrollo y la evaluación de modelos de procesamiento de lenguaje automatizado. Sin embargo, la creación y utilización de corpus conllevan una serie de desafíos legales y éticos, así como un conjunto de tareas de preprocesamiento de datos necesarias para garantizar la calidad y la representatividad de los conjuntos de datos resultantes.

En este capítulo, se abordarán tres dimensiones esenciales de la creación de corpus: en primer lugar, cuáles son las herramientas más comunes en la elaboración de corpus sintéticos, posteriormente se analizará el marco legal y ético que rige la recopilación y el uso de datos lingüísticos, destacando las consideraciones relacionadas con los derechos de autor, la privacidad de los usuarios y los acuerdos de licencia. Por último, se explorarán en detalle los principales recursos de preprocesamiento de datos, que incluyen, por ejemplo, la eliminación de ruido y la *tokenización* de palabras, con el propósito de preparar los datos para el posterior entrenamiento.

En resumen, a lo largo de este capítulo, se explorarán los desafíos y las consideraciones clave que los investigadores deben tener en cuenta al embarcarse en la creación de corpus sintéticos, con el fin de garantizar la integridad, la validez y la ética en la investigación lingüística.

1. Herramientas para la creación de corpus sintéticos

A pesar de la aparición del nuevo paradigma de los grandes modelos lingüísticos y los transformadores generativos preentrenados, en el ámbito de las lenguas con pocos recursos, la traducción automática neuronal (TAN) sigue siendo el enfoque más utilizado. Con todo, a pesar de que la TAN necesita menos datos bilingües para ser entrenada que los grandes modelos lingüísticos, disponer de esta cifra de millones de segmentos bilingües sigue siendo un desafío para las lenguas con pocos recursos, ya que el éxito del

motor depende directamente de esto. Por ello, en los últimos años se ha puesto el foco en el desarrollo de distintas estrategias para intentar encontrar una solución a la escasez de datos. En este sentido, podemos hablar de:

- TAN multilingüe (Lakew *et al.*, 2018), es decir, sistemas de TAN que son capaces de traducir entre múltiples idiomas, lo que significa que pueden trabajar con una variedad de combinaciones de idiomas.
- Transferencia de aprendizaje (Zoph *et al.*, 2016), que implica emplear el conocimiento aprendido en la traducción de un par de idiomas para mejorar la traducción en otros idiomas.
- Uso de lenguas próximas (Goyal *et al.*, 2016), que se centra en aprovechar las similitudes lingüísticas entre idiomas cercanos para mejorar la calidad de la traducción.
- TAN multimodal (Chowdhury *et al.*, 2018), que se refiere a sistemas de TAN que pueden manejar diferentes tipos de medios, como texto, imágenes y audio, para realizar traducciones más completas y contextuales.
- *Zero-shot* TAN (Currey *et al.*, 2019), por ejemplo, pivotaje de lenguas. Esta técnica busca traducir entre pares de idiomas para los cuales el sistema no ha sido entrenado directamente, a menudo utilizando un idioma intermedio como puente en el proceso de traducción.
- Aumento de datos (Fadaee *et al.*, 2017), que desglosaremos en los siguientes apartados.
- Bases de datos pseudo paralelas filtradas (Imankulova *et al.*, 2017). Este enfoque se refiere a la creación de conjuntos de datos que contienen pares de oraciones en diferentes idiomas, pero que no son traducciones literales la una de la otra.
- Meta-aprendizaje (Gu *et al.*, 2018), que se refiere a la capacidad de un sistema de TAN para aprender cómo aprender, es decir, adaptarse y mejorar su rendimiento en la traducción a medida que se le presentan nuevos desafíos o idiomas.

La mayoría de estas estrategias se pueden agrupar en dos grupos principales: aumentar la cantidad de datos aptos para el entrenamiento o modificar las técnicas de modelado para sacar más provecho de una cantidad menor de datos. En el caso particular de la investigación que nos ocupa, nos centraremos en el primer grupo.

1.1. Estrategias para el aumento de datos

Ranathunga *et al.* (2023) definen aumento de datos como el conjunto de técnicas usadas para crear datos adicionales bien mediante la modificación de los datos existentes o bien añadiendo datos de diferentes fuentes y que están destinadas a generar pares sintéticos de oraciones paralelas. Es la única de las estrategias mencionadas que no modifica la arquitectura neuronal. Los autores también dividen las estrategias para el aumento de datos en tres grandes grupos: reemplazo de oraciones o palabras para aumentar datos, *back-translation* y minería de corpus paralelos.

1.1.1. Reemplazo de oraciones y palabras para aumentar datos

Esta estrategia consiste en reemplazar las palabras o las oraciones de un subconjunto de oraciones del corpus paralelo o del corpus monolingüe. En el caso de reemplazar palabras, habitualmente se realiza con la ayuda de un diccionario bilingüe y se pueden seleccionar todas las palabras o palabras raras o poco frecuentes. Si bien, existe el riesgo de pérdida de fluidez al ejecutar esta técnica. Alternativamente, en lugar de reemplazar palabras, se puede efectuar el mismo proceso, pero reemplazando oraciones, siempre sin cambiar el sentido de la frase. Podemos ver ejemplos de esta técnica en Fadaee *et al.* (2018) y Liu *et al.* (2021).

1.1.2. Back-translation

La traducción inversa o *back-translation* es el proceso de traducir un corpus monolingüe en la lengua de llegada a la lengua origen, es decir, en la dirección contraria de traducción usando un motor de traducción automática existente. A partir de esta traducción, se genera un corpus paralelo sintético que se utiliza para entrenar el motor. Habitualmente, se emplea un corpus monolingüe en la lengua de llegada para mejorar la fluidez del motor de traducción automática, ya que se ha demostrado que hacerlo desde la lengua origen es contraproducente (Sennrich *et al.*, 2016a).

Sin embargo, esta técnica no está exenta de problemas. Por un lado, entrenar con esta técnica produce más ruido que con un corpus paralelo original, especialmente, si no se disponen de suficientes datos originales. Por eso, las investigaciones en traducción inversa se han centrado en corregir estos efectos mediante la selección de datos, el filtrado de datos, o el etiquetado de datos entre otros.

Como veremos en los siguientes capítulos relativos a la metodología y al proceso de entrenamiento, nos centraremos especialmente en las estrategias de selección de datos y en el etiquetado de los corpus.

- *Selección de datos*

Dentro de la traducción inversa, es necesario seleccionar adecuadamente el tipo de datos monolingües que se van a traducir. Traducir de manera global todos los datos monolingües a nuestra disposición no va a ser efectivo dado que el motor no sería capaz de seguir aprendiendo al darle demasiado ruido. En su lugar, se debería seleccionar un conjunto de datos que vaya en consonancia con el propósito del motor (Fadaee *et al.*, 2018; Lohar *et al.*, 2019). Adicionalmente, aunque se seleccionen los datos que se van a traducir a la inversa, el corpus resultante puede contener errores y ruido. Por esta razón, es recomendable también filtrar y limpiar datos (ver siguiente apartado).

- *Etiquetado del corpus*

Es posible que, a pesar de seleccionar y filtrar datos para generar un corpus paralelo sintético, este sea de menor calidad que un corpus paralelo original. Por esta razón, es recomendable identificar cada uno de los corpus o asignar cierto peso o importancia a las oraciones de cada uno de los corpus. En el apartado de metodología, mostraremos cómo combinamos ambos corpus para el entrenamiento del motor.

1.2. Minería de corpus paralelos

Esta estrategia consiste en buscas corpus comparables, es decir, texto de la misma temática que no es una traducción directa del otro, sino que contiene fragmentos que podrían ser equivalentes. Por ejemplo, es bastante frecuente recurrir a artículos de Wikipedia, artículos de noticias o a *web-crawling* masivo (Koehn, 2005). Para realizar esta herramienta, la técnica más usada es la generación de incrustaciones multilingües, es decir, se usa un modelo para aprender la representación de una oración en dos o más lenguas. Posteriormente, para cada una de esas oraciones en cada una de las lenguas se identifica a sus vecinos más cercanos como oraciones paralelas en la otra lengua. Actualmente, se han realizado muchas investigaciones utilizando modelos preentrenados como BERT multilingüe (mBERT) o XLM-R (Zhang *et al.*, 2020).

Una de las ventajas de aplicar esta técnica es que te permite entrenar motores en combinaciones de idiomas en las que no existen corpus bilingües, por ejemplo, cuando el inglés no está entre los pares de trabajo. Sin embargo, esta estrategia no está exenta de

peligros: es necesario efectuar un buen filtrado de los datos para evitar entrenar al motor con demasiado ruido, algo a lo que la TAN es especialmente sensible.

1.3. Consideraciones finales

Tras lo expuesto en los apartados anteriores, parece oportuno plantearse entrenar un sistema de TAN para un género textual específico y optimizar los recursos a partir de crear tus propios datos de entrenamiento. Para ello también hay que dilucidar hasta qué punto es lícito obtener datos de terceros con este propósito. Por eso la sección siguiente se plantea el marco legal para la creación de corpus con fines de investigación. Una vez revisadas las cuestiones legales, se detallan las acciones de preprocesamiento necesarias para el aumento de datos (apartado 3).

2. Marco legal para la creación de corpus

La creación de corpus, especialmente aquellos destinados a la investigación y desarrollo en el campo de la lingüística, la inteligencia artificial y la traducción automática, es un proceso esencial que requiere una cuidadosa consideración de aspectos legales y éticos. La obtención y el uso de datos lingüísticos y textuales plantean desafíos fundamentales en cuanto a derechos de autor, privacidad y responsabilidad ética. En este contexto, es crucial establecer un marco legal sólido que permita tanto la recopilación como el empleo de corpus de manera responsable y respetuosa. En este apartado, exploraremos las dimensiones legales y éticas que rodean la creación de corpus, analizando los pasos hacia la definición de un marco legal y ético y presentando consideraciones finales sobre la importancia de estos aspectos en la investigación y desarrollo lingüístico.

2.1. Hacia un marco legal y ético

Desde los primeros estudios en lingüística computacional hace 50 años, se hizo patente la necesidad de disponer de recursos lingüísticos (RL) como glosarios, tesauros, terminología y corpus. Dicha necesidad sigue estando patente en la actualidad, si cabe, en mayor medida tras los últimos avances tecnológicos en traducción automática. Por esta razón, diversos autores han expresado la necesidad de introducir las corrientes de ética en las tecnologías de la traducción (Morkeens *et al.*, 2020). Tampoco debemos olvidarnos en este apartado por el respeto a la legalidad europea y a la inteligencia artificial ética.

En un primer momento, los investigadores crearon estos recursos para su propio uso. Sin embargo, para avanzar en esta disciplina y en sus distintas ramas, era necesario compartir, distribuir, utilizar, reutilizar, ampliar y enriquecer los recursos. Tal y como explican Cieri & DiPersio (2015), el primer objetivo era convencer a los proveedores de datos que los investigadores lingüísticos y los desarrolladores de tecnología del lenguaje usaban sus datos para un buen propósito, es decir, no competitivo y para expandir el progreso científico. La mayoría de los creadores permitían la utilización de sus recursos para fines educativos o de investigación, sin coste o por un coste mínimo, especialmente si estaban financiados con dinero público mediante un acuerdo de pares, es decir, del centro de datos al usuario. Esta situación cambió drásticamente cuando el desarrollo de la tecnología del lenguaje supuso la aparición de negocios rentables y con ellos la idea de que los recursos podían ser una fuente de ingresos. Como consecuencia, muchos creadores de RL optaron por comercializar sus recursos (o dichos derechos) y, por tanto, haciendo indispensable una licencia para su distribución. Asimismo, Cieri & DiPersio (2015) también apuntan otros desafíos al modelo mencionado anteriormente:

- Recursos que están, y que se procesan, en la web.
- Muchos usuarios potenciales (comunidad virtual).
- Las fases de procesamiento pueden involucrar numerosas plataformas y herramientas.
- Necesidad de procesar licencias/permisos al momento.

Existen otros factores que también influyen a la hora de hablar de licencias. Por un lado, conviene recordar que nos encontramos con múltiples partes interesadas (los operadores de recursos que se encargan de hospedar el repositorio y mantener el software; los proveedores del servicio que acceden a los datos y al software, y los usuarios), lo que hace que la responsabilidad esté repartida, ya que un único ente no controla el recurso. Tampoco existe homogeneidad en las configuraciones de servicios web, repositorios de software, nubes y clústeres.

Labropoulou, Piperidis & Margoni (2016: 61) en su trabajo sobre los problemas legales para la interoperabilidad del proyecto OpenMinTeD identifican claramente cuáles son las figuras legales que el investigador debe tener en cuenta:

> *Copyright and the Sui Generis Database Right (SGDR) are the most relevant rights for TDM purposes (De Wolf & Partners, 2014; Guibault & Wiebe, 2013). Other rights or regulations such*

> *as personal data protection and Public Sector Information (PSI) may also play a role, sometimes an important one. (Keller et al., 2014). However, generally speaking these forms of legal regulation cannot be managed through a licensing approach and will therefore be addressed only to the extent that they are relevant in relation to the interoperability considerations covered in this paper.*
>
> *In accordance to the above, it is at the level of copyright licences for content and software and to the Terms of Use employed for services that we need to direct our analysis.*

Por tanto, tal y como expresan Rodríguez-Doncel & Labropoulou (2015: 49), en lo relativo a las licencias para RL, hay que considerar dos dimensiones:

> *(a) the license itself, either in the form of a proper legal document, or some loosely expressed legal notice, (b) the clear indication of the licensing terms on the LR, in the form of free text or conventional metadata.*

Los RL se consideran obras bajo propiedad intelectual y, como tales, están protegidos por las leyes de derechos de autor, esto es, no se pueden usar en contra de los términos establecidos por los autores. Sin embargo, las aproximaciones al comportamiento del usuario incluyendo los derechos de propiedad intelectual son muy dispares. Morkeens *et al.* (2020: 478) concluyen que:

> *At present, the attribution of translation copyright (and the reuse of translation as data) is subject to a number of conflicting and inconsistently-interpreted laws and conventions and thus remains somewhat unclear.*

Aun así, son los términos de uso los que establecen las acciones que se autorizan (modificación, redistribución, etc.) y las condiciones que se aplican (pago de un coste, obligación de atribución, etc.). Estos términos suelen incluirse en la documentación de los RL, pero no existe una homogeneización en la forma en la que se presentan: pueden aparecer en un documento de texto, dentro de una estructura de metadatos o simplemente mediante una mera mención a alguna licencia conocida (Open Data Commons o Creative Commons).

Rodríguez-Doncel & Labropoulou (2015: 54-56) identificaron las licencias más usadas en RL junto con su representación (véase Ilustración 24 de la página siguiente):

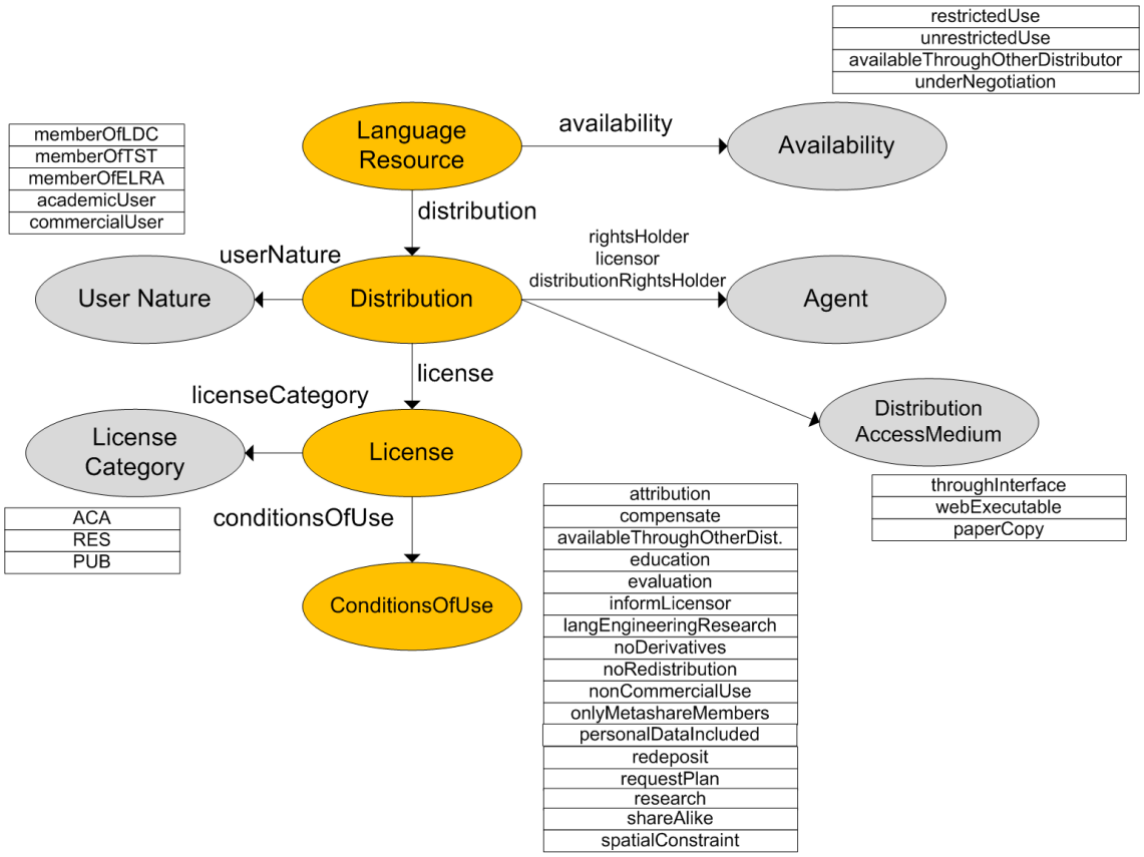


Ilustración 24. Licencias más usadas en RL según Rodríguez-Doncel & Labropoulou (2015: 54-56).

Estas son las licencias que listan los autores:

- Licencias tipo CC o FOSS que no requieren que el usuario firme ningún documento. Se trata de documentos legales con un texto general en el que se explican las condiciones de uso para los usuarios finales.
- Licencias estándar que incluyen elementos de ejemplo. Son documentos legales que ambas partes deben firman y el que se registran también los términos específicos para cada RL así como el coste que se debe pagar, en dónde se va a usar el RL, etc.
- Plantillas de licencia con términos extra potenciales. Hablamos de documentos legales que incluyen un texto general y unos términos potenciales (atribución, petición de un plan de investigación, uso de los recursos en una localización especial, etc.).
- Licencias no estándar como las licencias propias, avisos legales o los términos de uso.

Existen numerosas iniciativas para desarrollar un esquema común interoperativo (The European Language Grid[11], META-SHARE[12], PANACEA[13], LinguaGrid (Bosca *et al.*, 2012), CLARIN[14], LAPPS (Ide *et al.*, 2014), etc.). Sin embargo, para los objetivos de este trabajo, resulta mucho más interesante identificar las dimensiones de las licencias. Por tanto, utilizaremos el esquema de Cieri & DiPersio (2015) para garantizar la generación ética y legal de nuestro corpus, tal y como veremos en el capítulo VI:

- ¿De quién es la propiedad?
    - Usuario
    - Dominio público
    - Copyright
    - Limitada (contrato)
- ¿Cuáles son los límites de uso y para compartir?
    - No comercial
    - No trabajos derivados
    - Posibilidad de compartir
    - Atribución
    - No distribución
    - Otros
- ¿El usuario dispone de licencia?
- ¿Qué tipo de uso está permitido?
    - Educativo
    - Investigación
    - Desarrollo de tecnología (comercial)
    - Reventa
- ¿Qué tipo de procesamiento está permitido?
    - Creación de un trabajo derivativo (p. ej. transcripción)
    - Creación de un trabajo transformativo (p. ej. lista de frecuencias o modelo de lengua)

[11] Disponible en: https://live.european-language-grid.eu/. Fecha de consulta: 2 de septiembre de 2023.
[12] Disponible en: http://www.meta-net.eu/. Fecha de consulta: 15 de agosto de 2023
[13] PANACEA - Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language. Disponible en: http://www.panacea-lr.eu/. Fecha de consulta: 15 de agosto de 2023
[14] Disponible en: https://www.clarin.eu/portal. Fecha de consulta: 15 de agosto de 2023

2.2. Consideraciones finales

Como hemos visto en este apartado, el reaprovechamiento de datos textuales para el entrenamiento de motores de traducción automática comporta consideraciones legales y éticas. A lo largo de este apartado, se ha mostrado luz sobre las licencias y leyes en vigor que aplican a los RL, y que, por tanto, afectan a nuestra investigación. También se han referenciado las distintas iniciativas, especialmente desde el ámbito europeo, para compartir y contribuir con RL.

3. El preprocesamiento de los datos de entrenamiento de sistemas TAN

Una establecidos los límites legales y éticos de la generación de datos (apartado 2), en este apartado, nos centraremos en los recursos y pasos necesarios para el preprocesamiento de los datos de entrenamiento para los motores de TAN en el contexto de lenguas con pocos recursos. Debido a la poca disponibilidad de datos y recursos lingüísticos de estas lenguas, un preprocesamiento adecuado es fundamentales para mejorar la calidad de la traducción y abordar las características específicas de las lenguas con pocos recursos.

Como ya hemos visto en los apartados anteriores, el primer paso para entrenar un sistema de TAN es contar con un corpus paralelo adecuado. En el caso de las lenguas con pocos recursos, la disponibilidad de corpus paralelos puede ser limitada. En este sentido, es fundamental recopilar y crear corpus paralelos específicos para estas lenguas. Esto implica la colaboración de lingüistas y expertos bilingües para recolectar muestras de texto alineadas en ambas lenguas, preferiblemente con una anotación de calidad lingüística. Sin embargo, antes de entrenar cualquier motor, es necesario realizar una serie de pasos previos. A continuación, describiremos cada uno de los pasos previos necesarios para preprocesar el corpus de entrenamiento.

3.1. Eliminación de ruido

3.1.1. Normalización lingüística

Las lenguas con pocos recursos a menudo presentan desafíos adicionales debido a la falta de normas ortográficas o gramaticales establecidas. Por lo tanto, la normalización

lingüística es un paso importante en el preprocesamiento de la TAN. Esto implica corregir errores ortográficos y gramaticales, estandarizar la escritura y aplicar técnicas de limpieza de texto para mejorar la calidad de los datos de entrenamiento, como se detalla en la sección siguiente.

3.1.2. Limpieza del corpus

Habitualmente es necesario realizar una serie de comprobaciones antes de entrenar el motor para garantizar que nuestro corpus paralelo no cuente con:

- Entidades html/xml en lugar de su correspondiente carácter, por ejemplo, en lugar de «ñ» podríamos encontrarnos «ñ».
- Segmentos sin correspondencia con la otra lengua.
- Segmentos en lenguas que no sean las lenguas del corpus.
- Un porcentaje alto de expresiones numéricas en un mismo segmento.
- Etiquetas XML mezcladas con el texto.

3.1.3. Manejo de datos raros

Las lenguas con pocos recursos a menudo tienen palabras o estructuras gramaticales que son poco frecuentes en otros idiomas más comunes. Para mejorar la capacidad del modelo de TAN para manejar estos datos raros, es posible aplicar técnicas de aumento de datos como la sustitución de palabras poco frecuentes por sinónimos o la generación de palabras similares.

3.2. Fases del preprocesamiento

3.2.1. Segmentación de palabras o *tokenización*

En muchos idiomas, las palabras se escriben sin espacios entre ellas, lo que dificulta la segmentación adecuada del texto en unidades lingüísticas significativas. Para abordar este problema, se requiere una segmentación de palabras precisa antes de entrenar el modelo de NMT. Cabe resaltar que este proceso depende de las características de la lengua, ya que, por ejemplo, resulta útil separar palabras y signos de puntuación en español, pero se requiere de un *script* especial para hacerlo en un idioma como el chino donde no existe la separación entre palabras y hay que evitar ambigüedades.

Lo habitual es recurrir a un *tokenizador* ya existente que contemple la posibilidad de especificar la lengua como, por ejemplo, spaCY (Explotion AI, 2016), Standfor Core NLP (Manning *et al.*, 2014) o Moses (Moses, 2017).

Así funcionaría el *tokenizador* con la siguiente frase en inglés:

```
They're the teacher's best students.

They 're the teacher 's best students .
```

Ilustración 25. Ejemplo de *tokenización*

3.2.2. Truecasing

El *truecasing* es un proceso de NLP que se utiliza para convertir el texto a la forma correcta en letras mayúsculas o minúsculas de acuerdo con las reglas gramaticales y lingüísticas apropiadas de un idioma. Esta técnica es especialmente útil para normalizar el texto, especialmente si los textos tienen letras mayúsculas aleatorias o están completamente en mayúsculas. Antes de realizar esta operación, debemos contar con un modelo específicamente entrenado para *truecasing* a partir de diccionarios de formas de palabras y de los propios corpus.

Es importante recordar que el *truecasing* solo tiene sentido para aquellas lenguas que disponen de mayúsculas y minúsculas. Solo las lenguas que utilizan los siguientes alfabetos tienen letras mayúsculas y minúsculas:

- armenio
- cirílico
- griego
- latín

A continuación, podemos ver un ejemplo de *truecasing* en una frase en español ya *tokenizada*:

```
La Primavera llegará pronto a Barcelona .
```

Ilustración 26. Ejemplo de frase *tokenizada*

Después de aplicar el *script* para *truecasing* obtendríamos:

```
la primavera llegará pronto a Barcelona .
```

Ilustración 27. Ejemplo de frase *tokenizada* después de aplicar *truecasing*

3.2.3. Tratamiento de expresiones numéricas

Las expresiones numéricas requieren de una atención especial. Conviene recordar que existe un número infinito de expresiones numéricas diferentes, por lo que aprender la traducción de cada una de ellas sería imposible. La estrategia habitual para lidiar con expresiones numéricas es separar las cifras de las expresiones numéricas con espacios en blanco en el corpus de entrenamiento y en las oraciones que se van a traducir para posteriormente eliminar dichos espacios en la traducción.

3.2.4. Tratamiento de emails y URL

Estos dos elementos, emails y *URL*, en general no son traducibles y requieren también un tratamiento especial. Normalmente se reemplazan por unos códigos tanto en el corpus de entrenamiento como en la oración que se va a traducir y, posteriormente, se restauran una vez traducida la oración por los correspondientes originales:

- En el caso de los emails, lo habitual es reemplazarlos por el código @EMAIL@.
- En el caso de las *URL*, lo habitual es reemplazarlas por el código @URL@.

3.2.5. Subpalabras

Los primeros sistemas neuronales tenían un problema importante en la extensión del vocabulario, es decir, en el número de palabras diferentes (en realidad, de formas, es decir, de palabras flexionadas) capaz de traducir. Dado que cada palabra (en realidad forma) se convierte en un vector y el sistema trabaja con estos vectores, el número de palabras era finito. Si el número de palabras posible era muy grande, el sistema requería mucha memoria para trabajar, y si era pequeño, el número de palabras desconocidas era demasiado grande.

Con estos sistemas, ocurría a menudo que una oración como *I spent my holidays in Spain* en un sistema que no tenía *Spain* en su diccionario pero sí *France*, por ejemplo, se traducía por «He pasado mis vacaciones en Francia»[15]. Una estrategia utilizada era substituir las palabras desconocidas por <unk>, de manera que la oración de partida quedaba *I spent*

[15] Los ejemplos son propios.

my holidays in <unk>, que se traducía por «He pasado mis vacaciones en <unk>». Una vez obtenida la traducción se recuperaba el original de la palabra desconocida y se ponía en la traducción, o bien directamente, quedando «He pasado mis vacaciones en Spain», o bien consultando un diccionario, que si contenía el par *Spain*-España, resultaba en la traducción correcta: «He pasado mis vacaciones en España».

Desde hace unos pocos años la estrategia más extendida para solucionar el problema de la limitación de vocabulario es trabajar con subpalabras (*subwords*, en inglés), es decir, con fragmentos de palabras. De esta manera, las formas más frecuentes se utilizan enteras, pero las menos frecuentes se dividen en trozos que sean frecuentes. De esta manera se pueden representar todas las palabras, ya sean enteras o por fragmentos. El sistema traduce palabras y sub-palabras y dando como resultado una traducción con palabras y sub-palabras. Al final del proceso, se unen los trozos de las subpalabras traducidas.

Hoy en día existen los algoritmos más usados son: *subword-nmt* y *sentencepiece*. Como veremos un poco más adelante, *sentencepiece* va más allá del cálculo y uso de sub-palabras y pretende ser un *tokenizador* y *detokenizador* universal no supervisado.

- *Subword-nmt*

El estándar para este proceso es el creado por (Sennrich *et al.*, 2016) denominado *Byte Pair Encoding* (BPE). BPE es un compresor de datos que sustituye pares más frecuentes de bytes consecutivos por un byte que no existe en los datos.

Las unidades sub-palabra se calculan:

- Para lenguas que utilizan el mismo alfabeto: a la vez (es decir, concatenando los corpus) para la lengua de partida y de llegada.
- Para lenguas que no utilizan el mismo alfabeto: por separado para cada lengua.

- *Sentencepiece*

Sentencepiece es un *tokenizador* y *detokenizador* no supervisado y puede utilizarse como algoritmo para calcular y aplicar subpalabras y como procesador único del corpus, sin requerir ni *tokenización* ni *truecasing* previo, ya que los autores (Kudo & Richardson, 2018) declaran que «*SentencePiece allows us to make a purely end-to-end system that does not depend on language-specific pre/postprocessing*».

Por lo general, si los textos que se van a traducir son muy libres en el uso de oraciones en mayúsculas, es mejor optar por *tokenización*, *truecasing* y *sentencepiece*. Por el contrario, en los casos en que el uso de palabras u oraciones en mayúsculas están limitados a pocos casos repetitivos es mejor utilizar solo *sentencepiece*.

3.2.6. División del corpus en corpus de entrenamiento, corpus de validación y corpus de evaluación

Una vez el corpus paralelo está preprocesado siguiendo los criterios descritos en las secciones anteriores, es necesario dividir ese corpus en tres, de la siguiente forma:

- *train*: será el fragmento que sirva para entrenar el sistema.
- *val*: será el fragmento que se utilizará en cada iteración de entrenamiento.
- *eval*: será el fragmento que se utilice para evaluar el sistema una vez entrenado.

Habitualmente, se suele realizar esta tarea con *scripts* específicos dependiendo del número de oraciones finales del corpus, tal y como veremos en el siguiente capítulo.

CAPÍTULO V: Mejora y evaluación de la traducción automática

En la búsqueda constante por mejorar la calidad y precisión de la traducción automática, se ha convertido en una prioridad explorar dos aspectos fundamentales: el posprocesamiento de los resultados obtenidos a través de sistemas de traducción automática neuronal (TAN) y la evaluación de su desempeño. En este capítulo, se abordarán estas dos dimensiones esenciales en la optimización de la traducción automática.

En la primera sección de este capítulo, nos enfocaremos en la discusión de los recursos disponibles para el posprocesamiento de las traducciones automáticas. Exploraremos herramientas de edición y corrección, así como estrategias de adaptación de contenido que permiten personalizar las traducciones según el contexto y los requisitos del usuario final.

La segunda sección de este capítulo se centrará en los instrumentos de evaluación métrica, que son esenciales para medir la calidad y el rendimiento de los sistemas de TAN. La evaluación es un paso crítico en el ciclo de desarrollo de estos sistemas, y existen enfoques tanto automáticos como humanos para llevar a cabo esta tarea. Discutiremos las métricas automáticas, que incluyen medidas como BLEU, METEOR y TER, que proporcionan una evaluación cuantitativa rápida, pero a menudo limitada. También exploraremos la evaluación humana, que implica la participación de evaluadores humanos para proporcionar una perspectiva más rica y contextual sobre la calidad de las traducciones.

En resumen, este capítulo se sumerge en dos aspectos fundamentales en la mejora y evaluación de la traducción automática: el posprocesamiento de las traducciones generadas por sistemas de TAN y la utilización de instrumentos de evaluación métrica, tanto automáticos como humanos. A medida que avanzamos en la exploración de estos temas, esperamos arrojar luz sobre las estrategias y herramientas disponibles para perfeccionar las traducciones automáticas y evaluar su eficacia de manera efectiva.

1. Recursos para el posprocesamiento del resultado de la TAN

Aunque los sistemas de TAN han mejorado significativamente la calidad de las traducciones automáticas, todavía existen desafíos en cuanto a la fluidez, la coherencia y la corrección gramatical de las traducciones generadas. En este sentido, el posprocesamiento se ha convertido en una etapa crucial para mejorar la calidad de las traducciones automáticas. Este apartado tiene como objetivo analizar y describir las estrategias y recursos utilizados en el posprocesamiento de la traducción automática.

En primer lugar, podemos destacar el uso de expresiones regulares para corregir errores habituales. A pesar del preprocesamiento hecho y del corpus empleado para el entrenamiento del motor, la TAN es una tecnología imperfecta. Por este motivo, algunos investigadores han desarrollado pequeños *scripts* para corregir de manera automática determinado tipo de errores que siempre comete el motor. Es el caso de Guzmán (2007) que describe cómo se pueden emplear expresiones regulares para arreglar errores de ortografía, falta de concordancia de género y número, sintaxis errónea, redundancias y problemas de estilo. También vemos esta misma estrategia en Park *et al.* (2020) y Stymne (2011). Los primeros optan por realizar una posedición automática (APE, por sus siglas en inglés) del coreano y la segunda que utiliza sugerencias de corrección a partir de un corrector ortográfico para mejorar la concordancia y el orden de las palabras en alemán.

En segundo lugar, podemos hablar de la adaptación al dominio. Los sistemas de traducción automática generalmente están entrenados en datos de dominio amplio, lo que puede generar traducciones menos precisas y adecuadas para dominios específicos, como la medicina o la tecnología. En este sentido, se aplican técnicas de adaptación para mejorar la coherencia y la terminología en el texto traducido, utilizando recursos como glosarios, memorias de traducción y corpus específicos del dominio. Habitualmente se suele realizar mediante reemplazos automáticos a partir de una base de datos terminológica. La mayoría de los sistemas neuronales comerciales más utilizados hoy en día ya ofrecen la posibilidad de importar un glosario unívoco en el que se establezca un término fuente A y un término de llegada A. El propio sistema está entrenado para ser

capaz de adaptar el género y el número de los términos, como por ejemplo, el sistema comercial *DeepL*[16].

Una vez ejecutados los pasos anteriores, otra estrategia de posprocesamiento se enfoca en la edición y corrección de las traducciones automáticas, también llamada posedición. Este enfoque implica la revisión y corrección manual de los errores gramaticales, la ortografía y la coherencia del texto traducido. Los traductores humanos o los editores especializados desempeñan un papel crucial en esta etapa, empleando su experiencia y conocimiento lingüístico para mejorar la calidad general de la traducción.

Por último, cabe destacar la realimentación del usuario. La interacción entre el usuario y el sistema de traducción automática puede ser aprovechada como una estrategia de posprocesamiento efectiva. Los usuarios pueden proporcionar comentarios y correcciones sobre las traducciones generadas, lo que permite al sistema aprender y mejorar continuamente. Este tipo de TA se denomina TA interactiva y la máquina aprende en tiempo real sobre las elecciones del usuario. Además, los sistemas de traducción automática pueden usar técnicas de aprendizaje activo para solicitar a los usuarios que validen o corrijan las traducciones, mejorando así la calidad del resultado final.

2. Instrumentos de evaluación métrica

En este apartado, se analizarán los principales instrumentos de evaluación métrica de la TAN, así como su utilidad y limitaciones. Si bien, antes de entrar dentro de los distintos tipos, conviene definir qué se entiende por calidad dentro del mundo de la traducción en general y, en particular, dentro de la traducción automática.

En un sentido amplio, la calidad en la traducción se refiere tanto al producto final, es decir, el material traducido, como al proceso, esto es, el servicio ofrecido (Gouadec, 2010:270). La noción de calidad es relativa (Grbić, 2008: 232), ya que en la mayoría de los casos depende del contexto en el que se hizo, de las expectativas del cliente y del propósito de la traducción. Esto ha dado lugar según Doherty, (2017: 131) a la aparición de definiciones implícitas y operacionalizadas de manera diferente junto con sus

[16] Disponible en: https://www.deepl.com/es/translator. Fecha de consulta: 20 de septiembre de 2023.

respectivas medidas. Por esta razón, nos inclinamos por la definición amplia que dan Koby *et al.* (2014: 416):

> *A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs.*

Al añadirle el paradigma de la traducción automática, la calidad normalmente se define como un medio para un fin, principalmente para mejorar los sistemas de TA, por lo que siempre ha prevalecido un enfoque pragmático en el que se combinan la evaluación humana y la automática (Doherty, 2017:133). Rossi *et al.* (2023: 53) declaran que este enfoque pragmático implica usar indicadores medible de usabilidad como índices de satisfacción de usuarios, aumentos de productividad durante la posedición o el incremento de las ventas a partir de la descripción de productos traducidos con un motor de TA. También se debe tener en cuenta que evaluar la calidad de una traducción no es una tarea sencilla, ya que intervienen diversos factores. Por un lado, habitualmente suele haber más de una traducción válida. Por otro lado, si la evaluación la realizan humanos, esta suele ser subjetiva, suele llevar más tiempo y necesitar más recursos. Por el contrario, si la evaluación se efectúa mediante algoritmos automáticos, puede ser menos eficaz, a pesar de ser más rápida y barata. Como veremos más abajo, ambos tipos de evaluación tienen ventajas y desventajas, por lo que la elección de un método u otro dependerá de las necesidades del proyecto.

Como hemos avanzado anteriormente, podemos clasificar los métodos de evaluación en dos grandes grupos: automáticos o humanos. Por métodos automáticos entendemos aquellos en los que no se da participación humana en el proceso de evaluación. Por el contrario, una evaluación humana requiere la intervención manual de los humanos en la fase de evaluación. A su vez, dividiremos la evaluación humana es dos grupos dependiendo de su la valoración de expresa directamente (*Directly Expressed Judgment, DEJ,* por sus siglas en inglés) o no (*Non-directly Expressed Judgment, non-DEJ,* por sus siglas en inglés) (Chatzikoumi, 2020: 3). Los métodos *DEJ* se basan en pedirle a una persona que establezca si la traducción es buena, aceptable o mala y los otros en pedirles a los evaluadores que clasifiquen o corrijan errores o que completen una tarea de relleno de huecos a partir de la compresión de un texto traducido automáticamente.

2.1. Métricas automáticas de evaluación de la calidad

Las métricas automáticas son aquellas en las que los humanos no participan directamente en el proceso de evaluación, únicamente para desarrollar el algoritmo o para ejecutarlo. Chatzikoumi, (2020:4) diferencia entre métricas que arrojan un valor de similitud con respecto a una traducción de referencia, métricas de confianza o de estimación de la calidad y evaluación de diagnóstico basada en puntos de control.

2.1.1. Métricas basadas en una traducción de referencia

Estas métricas arrojan un grado de similitud con respecto a una traducción humana de referencia, también llamada en inglés *gold standard* (Papineni *et al.*, 2002).

- *Distancia de edición*

Esta métrica se basa en la distancia de Levenshtein (Levenshtein, 1966). La distancia de edición entre la cadena traducida automáticamente y la cadena de referencia es el número mínimo de operaciones necesarias para convertir la cadena traducida automática en la cadena de referencia (Navarro, 2001). Habitualmente, se tienen en cuenta palabras enteras y no caracteres.

Dentro de esta categoría destaca Word Error Rate (WER) (Niessen, 2000) que es la proporción de la suma de las operaciones de edición en una cadena traducida automática con respecto al número de palabras de la cadena de referencia. Existen distintas variaciones de esta técnica como WERg (Blatz *et al.* 2004); Translation Edit Rate (TER) (Snover *et al.* 2006); Multiple ReferenceWER (MWER) (Niessen *et al.* 2000); inversión WER (invWER) (Leusch, Ueffing y Ney 2003); sentence error rate (SER) (Tomás, Mas y Casacuberta 2003); cover disjoint error rate (CDER) (Leusch, Ueffing y Ney 2006), y hybrid TER (HyTER) (Dreyer y Marcu 2012). Las principales diferencias entre estas variaciones y adaptaciones tienen que ver con el número de traducciones de referencia usadas y también en si consideran o no los movimientos de palabras o construcciones como una operación de edición.

- *Precisión y recuperación*

En el caso de este tipo de métricas, la precisión es la proporción entre los unigramas aceptables de la cadena traducida automáticamente y el número de unigramas de la misma cadena. Si la cadena se compone de 10 unigramas o palabras y seis de ellos son aceptables, entonces la precisión será de 6/10. Por su parte, recuperación es la proporción de

unigramas aceptables de la cadena traducida automáticamente con respecto al número de unigramas de la traducción de referencia, es decir, el número de unigramas deseado.

Dentro de este grupo, la métrica más utilizada hasta el momento es el *bilingual Evaluation Understudy* (BLEU) (Papineni *et al.*, 2002). En esta métrica la calidad de la traducción automática se mide de acuerdo con su semejanza a una traducción humana. El grado de similitud se expresa en una escala de 0 a 1, donde 0 es la total diferencia y 1 es la similitud total. A pesar de que es una métrica que todavía se emplea ampliamente tanto en el mundo académico como empresarial, algunos autores han propuesto variaciones a BLEU o métricas alternativas, buscando una mayor correlación de la métrica con valoraciones humanas. Este es el caso de SacreBLEU (Post, 2018) o NIST (Doddington 2002) que da más relevancia a los unigramas informativos o COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei *et al.*, 2020). COMET es definido por sus autores como un marco basado en PyTorch que permite entrenar modelos de evaluación de traducción automática multilingües y adaptables que funcionan como métricas.

Por otra parte, también debemos mencionar en este grupo métricas que combinan precisión y recuperación, como la métrica *F-measure* (Melamed, Green y Turian 2003); *Character n-gram F-score* (CHRF) (Popovi´c 2015); *Metric for Evaluation of Translation with Explicit Ordering* (METEOR) (Banerjee y Lavie 2005); *ParaEval* (Zhou *et al.*, 2016), y la métrica *Length Penalty, Precision, n-gram Position difference Penalty and Recall* (LEPOR) (Han *et al.*, 2012).

2.1.2. Confianza o estimación de la calidad

La estimación de la calidad (*QE*, por sus siglas en inglés) se usa para predecir la calidad del resultado de un sistema para un input determinado sin información sobre el resultado deseado (Specia *et al.*, 2009). Por tanto, las métricas de estimación de la calidad no son métricas por sí mismas, sino que funciona como un proxy. Normalmente, la plataforma de QE se compone de dos módulos independientes: un módulo de extracción de características y un módulo de aprendizaje automático. El módulo de aprendizaje automático usa las características extraídas del texto de origen y de llegada para crear modelos de QE gracias al uso de algoritmos de regresión y clasificación. Las características extraídas se pueden agrupar en tres tipos: complejidad, fluidez y adecuación. La complejidad se refiere a la dificultad en la comprensión de una traducción, es decir, cuán fácil o difícil es entender el contenido traducido en comparación con el

texto original. La complejidad está relacionada con la elección de palabras, la estructura de las oraciones, la gramática y la sintaxis utilizadas en la traducción. Por su parte, la fluidez se refiere a la naturalidad y la suavidad de la traducción, esto es, cuán bien fluye el texto traducido y cuán coherente es en términos lingüísticos. Suele estar relacionada con aspectos como la coherencia léxica, la estructura de las oraciones, la elección de palabras y la ausencia de fragmentación o interrupciones en el texto traducido. Por último, la adecuación se refiere a la capacidad de la traducción para transmitir el significado y la intención del texto original, es decir, en qué medida la traducción cumple con el propósito de comunicación del texto de origen. La adecuación se relaciona con la fidelidad y la precisión de la traducción. Una traducción adecuada debe capturar el significado, el tono y la intención del texto original de manera efectiva. Puede involucrar la traducción de términos específicos, la interpretación de metáforas o la transmisión de connotaciones culturales.

2.1.3. Evaluación de diagnóstico basado en puntos de control

Este tipo de evaluación creada por Zhou *et al.* (2008) se basa en predefinir y extraer características a nivel lingüístico como palabras ambiguas o frases nominales o preposicionales a partir de un corpus paralelo para crear puntos de control. Estos puntos se usan para monitorizar la traducción de fenómenos lingüísticos de importancia y proporcionar un diagnóstico.

2.2. Métricas de evaluación humana

Las métricas de evaluación humana son aquellas en que las personas participan directamente en el proceso de evaluación. Como hemos visto anteriormente, las podemos subclasificar en métricas en las que la valoración se expresa directamente o no. A continuación, describiremos cada una de las subcategorías dentro de estos grandes grupos.

2.2.1. Métricas en las que la valoración se expresa directamente (*DEJ*)

Dentro de las métricas DEJ, podemos agrupar todas aquellas en las que los evaluadores expresan directamente la valoración de la calidad de la traducción. Habitualmente estas se pueden dividir en evaluaciones de adecuación y fluidez, posicionar en una clasificación distintos sistemas de TA o expresar directamente una valoración de la traducción (*DA*, en sus siglas en inglés).

- *Evaluación de adecuación y fluidez*

La Translation Automation User Society (TAUS) desarrolló en 2011 el Dynamic Quality Framework (DQF). Se trata de una plataforma que permite al usuario elegir si desea evaluar la adecuación, la fluidez o ambas. La evaluación de ambos aspectos suele ser lo habitual. La evaluación de pondera en una escala de cuatro puntos: la adecuación puede ser en la totalidad de los casos, en la mayoría de los casos, en algún caso o ningún caso y la fluidez puede ser total, buena, no fluido o incompresible en la lengua de llegada. Para que los resultados sean fiables, los evaluadores deben disponer de unos criterios muy claros en los que se establezca la diferencia entre los cuatro niveles y deben contar con la suficiente preparación para llevar a cabo la evaluación.

- *Ranking de sistemas de TA*

Este tipo de evaluación consiste en realizar una comparación para elegir el mejor de entre varios sistemas. No existe, un número máximo de motores para comparar. Algunos autores como Görög (2014) sugieren que se debe realizar una comparación de tres motores, ya que si se añaden más opciones a nivel de segmento el evaluador puede perder robustez en sus decisiones. No obstante, las campañas de evaluación de los grupos de trabajo sobre evaluación tanto estadística como neuronal optan por evaluar más cantidad de motores. Dentro del marco DQF, podemos encontrar dos tipos de ranking: un ranking del mejor sistema de TA o un ranking de la mejor traducción para el segmento evaluado.

- *Expresar directamente una valoración*

Este tipo de evaluación consiste en expresar la valoración de la calidad de la traducción automática en una escala de valoración continua (Graham *et al.* 2015), lo que refleja el grado en el que una traducción es mejor que otra. Actualmente, se orienta a la adecuación (Bentivogli *et al.* 2018).

2.2.2. Métricas en las que la valoración no se expresa directamente (*Non-DEJ*)

En este tipo de métricas, la valoración humana no se expresa de manera directa. Dentro de este grupo, podemos distinguir entre métricas semiautomáticas, la realización de tareas que requiere la comprensión del texto traducción de manera automática o la clasificación, análisis y corrección del resultado de la traducción automática.

- *Métricas semiautomáticas*

Las métricas semiautomáticas también denominadas evaluaciones en las que se pone al humano en el centro son variaciones de las métricas automáticas con la intervención de anotadores. Así tenemos HTER, HBLEU y HMETEOR (Snover *et al.* 2006). En la HTER, por ejemplo, además de tener el cálculo automático del número de ediciones necesarias para que la traducción automática sea idéntica a la traducción de referencia, un anotador crea una nueva referencia de traducción al hacer los menos cambios posibles en la traducción automática o en la traducción de referencia. Este tipo de métricas consiguen una mayor correlación con las valoraciones humanas, pero no son indicativas del esfuerzo del anotador.

- *Evaluaciones a partir de la realización de una tarea*

Una manera un poco diferente de evaluar la calidad de una traducción automática es mediante la evaluación de la eficacia de su resultado para realizar una tarea determinada. Podemos encontrar ejemplos de este tipo de tareas de evaluación al pedirle a los participantes que detecten la parte más relevante de información en un texto, que respondan preguntas sobre el contenido del texto o que completen los blancos en una traducción de referencia.

- *Análisis y clasificación de errores*

El análisis y clasificación de errores es uno de los tipos de evaluación humana más empleados. En este apartado nos vamos a centrar en el modelo harmonizado de Multidimensional Quality Metrics (MQM) y la tipología de errores DQF (Lommel *et al.*, 2014 y Görög, 2014) y la taxonomía de errores propuesta por Popovi´c (2018).

QTLaunchPad desarrolló el MQM, en el que se definen la calidad de la traducción humana y automática y se describe el proceso de evaluación bajo unos estándares específicos. Para ello, el proceso de evaluación se compone de detectar los parámetros o dimensiones de la traducción (lengua/local, dominio, terminología, tipo de texto, audiencia, propósito, registro, estilo, correspondencia del contenido, modalidad del resultado, formato del archivo y tecnología de producción), seleccionar las categorías de errores relevantes y anotar dichos errores. También se ofrece distintos cuatro niveles de gravedad del error: crítico, importante, menor y preferencial.

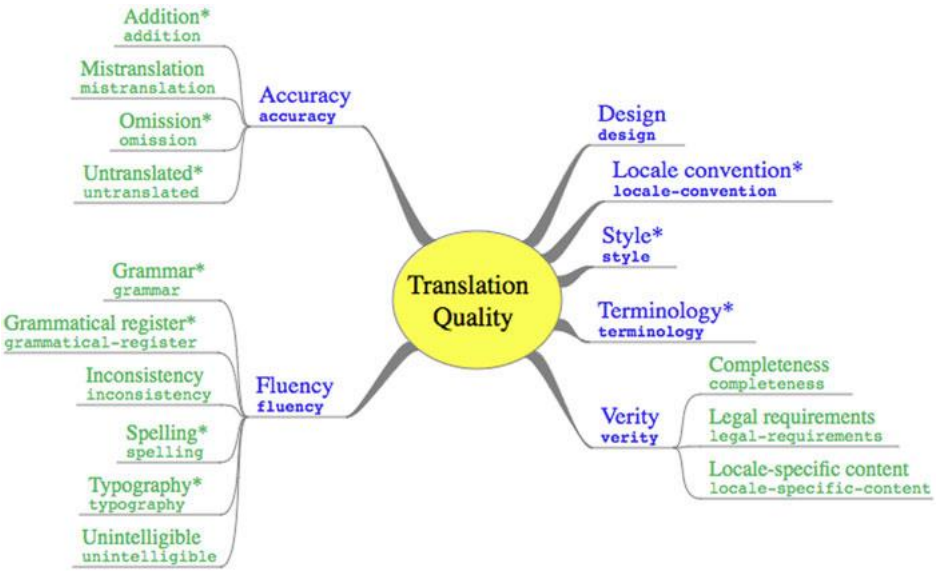


Ilustración 28. Errores principales en MQM.

En cuanto a la tipología de errores DQF, esta se compone de siete categorías: precisión, fluidez, terminología, estilo, convención local, diseño y veracidad. A su vez las categorías precisión y fluidez cuentan con subcategorías. En la categoría precisión, encontramos cinco subcategorías: adición, omisión, palabras no traducidas, concordancia exacta de la memoria de traducción impropia y traducción incorrecta. Por su parte, dentro de fluidez podemos dividir los errores en codificación de caracteres, ortografía, puntuación, enlace o referencia cruzada, registro gramatical, inconsistencia y gramática.

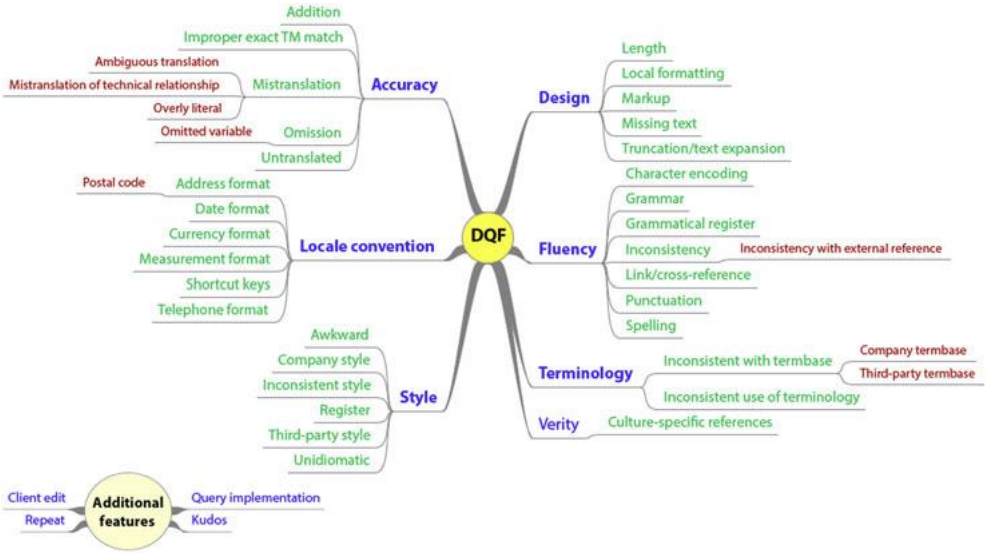


Ilustración 29. Subcategorías harmonizadas DQF-MQM.

Por su parte, Popovi´c (2018) sugiere una taxonomía general basada en un conjunto de categorías amplias que se pueden expandir en subniveles. Las categorías principales son léxico, morfología, sintaxis, semántica, ortografía y demasiados errores. En la siguiente tabla, se pueden apreciar los subniveles recogidos en Chatzikoumi, (2020: 17).

Level 1	Level 2	Level 3
Lexis	Mistranslation	Terminology
	Addition	
	Omission	
	Untranslated	
	Should not be translated	
Morphology	Inflection	Tense, number, person
		Case, number, gender
	Derivation	POS
		Verb aspect
	Composition	
Syntax	Word order	Range
	Phrase order	Range
Semantic	Multi-word expressions	
	Collocations	
	Disambiguation	
Orthography	Capitalisation	
	Punctuation	
	Spelling	
Too many errors		

Tabla 1. Taxonomía de errores de Popovi´c (2018)

Como podemos apreciar en la tabla anterior, esta taxonomía se utiliza para identificar y clasificar una amplia gama de errores en la traducción automática, lo que permite una evaluación más exhaustiva y precisa del rendimiento de los sistemas de traducción automática.

- *Posedición*

Sánchez y Rico (2020: 88) entienden la PE como «los cambios que se realizan sobre la traducción producida por un sistema de TA para que alcance unos niveles de calidad establecidos previamente» y admiten que la definición de PE está ligada a la evolución de la TA. En la misma línea, González y Rico (2021: 3) definen posedición como «la corrección del texto generado por una máquina para que cumpla con el nivel de calidad

exigido y puede aplicarse en distintos grados en función del uso final que se le vaya a dar al texto traducido». Por su parte, O'Brien (2023:106) define a la posedición como una tarea de procesamiento lingüístico bilingüe en la que se tiene que identificar errores, diseñar una estrategia para corregirlos e implementar esas revisiones. Sánchez-Gijón (2016: 152) también recoge una definición amplia de la posedición «en la que el traductor llega a dominar todos los procesos del proyecto de traducción con TA, más allá de la fase de traducción inicial en la que se validan las propuestas obtenidas automáticamente». Entre estos procesos, se encuentran la resolución de «problemas de adecuación, fluidez, naturalidad y sintonía del texto de llegada con un encargo de traducción determinado».

Habitualmente, podemos distinguir entre posedición completa o ligera. Massardo *et al.* (2016) definen la posedición completa como aquella le da una calidad humana al resultado del motor de TA. Por calidad humana entiende un texto que sea comprensible, preciso, estilísticamente adecuado, con una sintaxis normal y una gramática y puntuación correctas. Por el contrario, una posedición ligera es aquella en la que solo se arreglan los errores necesarios para que la traducción sea comprensible. Es aquí cuando hablamos de calidad suficiente o «*good enough*».

Habitualmente esta tarea es realizada por un traductor y se suele usar cómo métrica al medir el esfuerzo temporal y cognitivo del poseditor (Lacruz *et al.*, 2014).

TAUS estableció en 2010 unas instrucciones para poseditar que O'Brien recoge (2023: 109) en la siguiente tabla:

Light post-editing	Full post-editing
Aim for semantically correct translation.	Aim for grammatically, syntactically and semantically correct translation.
	Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms.
Ensure that no information has been accidentally added or omitted.	Ensure that no information has been accidentally added or omitted.
Edit any offensive, inappropriate or culturally unacceptable content.	Edit any offensive, inappropriate or culturally unacceptable content.
Use as much of the raw MT output as possible.	Use as much of the raw MT output as possible.
Basic rules regarding spelling apply.	Apply basic rules regarding spelling, punctuation and hyphenation.
No need to implement corrections that are of a stylistic nature only.	
No need to restructure sentences solely to improve the natural flow of the text.	
	Ensure that formatting is correct.

Tabla 2. Instrucciones de posedición de TAUS recogidas en O'Brien (2023: 109)

Como vemos, las instrucciones de TAUS invitan a reutilizar al máximo posible la traducción automática en bruto y asocian conceptualmente los dos niveles de posedición a niveles de calidad. La posedición ligera se asocia a una calidad meramente comprensible, mientras que la posedición completa se equipara a la calidad obtenida mediante una traducción humana.

2.3. Consideraciones finales

Como hemos visto en los apartados anteriores, la evaluación de la traducción automática se puede realizar mediante distintas técnicas. Teniendo en cuenta las ventajas y desventajas de cada una de ellas, lo ideal es combinar métricas automáticas y humanas para disponer de un método fiable. Cabe destacar, no obstante, que buena parte de las métricas antes mencionadas, especialmente las automáticas, miden el texto traducido, pero no la recepción de la traducción. Por otra parte, buena parte de las evaluaciones humanas buscan establecer la ausencia de errores y no necesariamente la adecuación a un contexto comunicativo concreto ni el mejor uso posible o uso natural de la lengua de llegada.

Por último, en lo que están de acuerdo todos los autores consultados es en que el mejor método de evaluación será aquel que se adapte a los objetivos del proyecto o de la traducción, tal y como explicaremos en el siguiente capítulo de metodología de esta investigación.

MARCO METODOLÓGICO

El presente bloque de capítulos constituye un elemento central en el desarrollo de esta tesis, ya que comprende no solo el marco metodológico empleado para alcanzar el objetivo principal de esta investigación: diseñar un protocolo de entrenamiento de un motor de traducción automática para las redes sociales dentro del contexto de las lenguas con menos recursos; sino también la propuesta de evaluación de este.

En el capítulo VI se proporciona una visión detallada de los métodos y procedimientos utilizados en la creación del corpus de entrenamiento. En primer lugar, se presenta una descripción general de los conjuntos de datos empleados y cómo se han obtenido, posteriormente se describe cómo se han preprocesado dichos textos.

El capítulo VII se centra en el entrenamiento y evaluación automática del motor de traducción automática. Primero se describe la tecnología empleada y posteriormente se evalúa automáticamente y se completa esta evaluación con un nuevo enfoque diseñado para extraer la percepción de los usuarios finales y validado mediante una prueba piloto descrita en el capítulo VIII. A través de una descripción minuciosa de las etapas y enfoques usados, se espera sentar las bases para un análisis riguroso y crítico de los resultados obtenidos, así como para la identificación de posibles mejoras.

Tras las conclusiones extraídas a partir del análisis de la prueba piloto, se aborda en el capítulo IX el nuevo procedimiento de entrenamiento realizado tras la introducción de mejoras.

Finalmente, en el capítulo X se describe la triple evaluación final del motor.

CAPÍTULO VI: Creación del corpus de entrenamiento ES-GL

El presente capítulo presenta el panorama completo del entrenamiento del motor de traducción automática del castellano al gallego en el ámbito de las redes sociales: la tecnología neuronal seleccionada, el proceso de recopilación de corpus, las estrategias aplicadas para aprovechar al máximo los datos disponibles y el proceso de entrenamiento en sí. Estos aspectos sientan las bases para comprender el desempeño y los resultados obtenidos, que serán discutidos en detalle en los siguientes capítulos de esta tesis.

El primer aspecto para destacar es la elección de la tecnología neuronal para entrenar el motor de traducción automática. En este sentido, se opta por emplear una arquitectura de red neuronal profunda, tipo *transformer* (véase CAPÍTULO III) conocida por su capacidad para modelar eficientemente las relaciones a largo plazo en el texto y capturar sutilezas lingüísticas. Además, se escogen plataformas o tecnología de libre acceso tal y como veremos más adelante.

En cuanto al corpus utilizado para el entrenamiento, se constata la poca disponibilidad de datos generales y la inexistencia de corpus propios del dominio de las redes sociales. Por esta razón, dado que el motor está orientado a un dominio específico y que crear un corpus único específico lo suficientemente grande como para poder entrenar no es viable, se opta por combinar datos genéricos con datos específicos, por lo que se lleva a cabo un minucioso proceso de recopilación de datos generales y se extraen datos de las redes sociales para esta investigación. Se recolectan muestras representativas de textos en castellano y gallego provenientes de diversas plataformas y se realizan tareas de limpieza y filtrado para garantizar la calidad y coherencia de los datos.

Para aprovechar al máximo los datos disponibles y optimizar el rendimiento del motor, se aplican diversas estrategias mencionadas en el marco teórico. Entre ellas, se emplea el aumento de datos mediante traducción inversa. Estas estrategias permiten aumentar la diversidad de los datos de entrenamiento y mejorar la capacidad del motor para lidiar con los diferentes estilos y registros presentes en las redes sociales.

Finalmente, se describe el proceso de entrenamiento del motor de traducción automática. Se detallan los parámetros utilizados, como el tamaño del lote, la tasa de aprendizaje y la función de pérdida, así como las técnicas de optimización empleadas para ajustar los pesos de la red neuronal. Asimismo, se aborda la duración y recursos computacionales necesarios para completar el proceso de entrenamiento.

1. Recopilación y creación de los corpus de entrenamiento

El primer paso necesario para entrenar un motor es saber de cuántos datos bilingües se dispone y de qué dominio son para determinar si son suficientes para un entrenamiento exitoso o si, por el contrario, se necesita recurrir a alguna de las estrategias de aumento de datos mencionadas en el marco teórico. Por un lado, gracias a la fotografía de la industria realizada en el CAPÍTULO I, sabemos que debemos contar con un corpus general de al menos entre 500 000 y 1 000 000 de segmentos bilingües. Por otro lado, revisamos los resultados de estudios previos en el ámbito de las redes sociales y exploramos los datos de este dominio disponibles en gallego.

1.1. Recopilación de corpus

En el momento de recopilación de los datos, en el año 2020, partimos de la investigación realizada previamente sobre recursos de procesamiento del lenguaje natural en gallego (do Campo & Sánchez-Gijón, 2019: 17-19). El único corpus bilingüe y con licencia Creative Commons 3.0 disponible en aquel entonces era una partición del Corpus Lingüístico da Universidade de Vigo (CLUVI). El corpus contiene 112 048 segmentos bilingües del ámbito principalmente administrativo. Necesitamos, por tanto, un corpus general mucho más amplio.

Posteriormente, en septiembre de 2021, la Comisión Europea mediante el programa CEF (por sus siglas en inglés: *Connecting Europe Facility*) publicó el corpus *Paracrawl* (Bañón *et al.*, 2020), creado a partir de páginas webs y disponible en los idiomas oficiales europeos y también en otros idiomas, como el gallego. La versión 9 de este corpus *Paracrawl* permite la descarga de un corpus de dominio general del castellano al gallego de 1 879 651 segmentos bilingües y 44 626 184 palabras bajo licencia CC0. CC0 es una licencia extremadamente permisiva que permite que el creador renuncie a todos sus

derechos de autor y derechos conexos sobre una obra. En otras palabras, el creador declara que no tiene ningún interés legal en la obra y la pone en el dominio público. Bajo CC0, cualquier persona puede utilizar, copiar, modificar, distribuir y realizar cualquier obra sin necesidad de obtener permiso o pagar regalías al creador original. Este es el corpus general que se usó en nuestro entrenamiento combinado con un corpus específico de Twitter del que hablaremos a continuación.

1.2. Creación del corpus específico

Ninguno de los recursos y corpus en gallego consultados hasta la fecha de redacción de esta tesis pertenece al dominio de las redes sociales. No obstante, sí encontramos un proyecto reciente de Ortega *et al*. (2023) en el que se crea un motor neuronal genérico de español a gallego, empleando una lengua cercana como es el portugués para obtener más datos. El corpus al que recurren es principalmente el corpus administrativo de la Unión Europea del castellano al portugués, que luego transliteran al gallego. También Oliver *et al*. (2023) están en proceso de creación de un motor multilingüe para las lenguas ibéricas: castellano, portugués, gallego, catalán, asturiano, aranés, aragonés y gascón. Aunque el proyecto está todavía en curso a la fecha de escritura de esta tesis, el objetivo es, por un lado, generar corpus nuevos especialmente para el aranés, aragonés y asturiano y, por el otro, utilizar un motor multilingüe. Sin embargo, en ambos casos, el dominio del motor es general y no específico de las redes sociales.

Si nos fijamos en la bibliografía relativa a investigaciones en redes sociales, se puede apreciar un considerable trabajo en análisis de sentimiento y traducción de campos de contenido generados por el usuario. Por ejemplo, Do Carmo *et al.* (2023) investigan si la TAN es capaz de traducir correctamente emociones en tuits. Concluyen que las métricas automáticas actuales de evaluación de la MT son incapaces de detectar esta particularidad lingüística de los tuits y que, por ello, se debe plantear una estrategia de evaluación que tenga en cuenta esta particularidad. Por su parte, Lohar *et al.* (2019) intentan comparar la calidad de tuits traducidos con traducción automática estadística y neuronal jugando con distintos tamaños y tipos de corpus de entrenamiento. Los resultados de esta investigación muestran que usar un corpus específico de redes sociales pequeño no es relevante para el entrenamiento neuronal, aunque sí se mejoran los resultados cuando se usa traducción inversa y se combina el corpus específico con un corpus genérico. Este procedimiento en concreto es el que empleamos en nuestra investigación: crear un corpus monolingüe en

gallego extraído de Twitter, traducirlo al castellano para generar un corpus bilingüe y combinarlo con el corpus genérico *Paracrawl* durante el entrenamiento del motor.

La licencia para utilizar datos de Twitter para fines de investigación es un aspecto crítico que se debe tener en cuenta previo al trabajo dentro de esta plataforma. Twitter proporciona acceso a su API (Interfaz de Programación de Aplicaciones) a investigadores y desarrolladores, lo que permite recopilar datos públicos de la plataforma para fines de investigación. Sin embargo, es importante tener en cuenta que Twitter tiene políticas y restricciones específicas en cuanto al uso de sus datos, que incluyen limitaciones en la redistribución de datos, la protección de la privacidad de los usuarios y la prohibición de ciertos tipos de actividades, como la recopilación de datos con fines de vigilancia. Los investigadores deben seguir cuidadosamente los términos y condiciones de la licencia de Twitter y respetar las restricciones establecidas para garantizar un uso ético y legal de los datos en sus investigaciones. Si seguimos el esquema de Cieri y DiPersio (2015) referenciado en el marco legal de esta tesis, podemos concluir que:

- ¿De quién es la propiedad?

Los datos de Twitter son propiedad de los usuarios que generan y comparten contenido en la plataforma. Estos datos son públicos en su mayoría, ya que los usuarios publican mensajes de forma voluntaria y pueden configurar la visibilidad de sus cuentas y tuits. Por eso, en las siguientes secciones listaremos siempre las cuentas de Twitter.

- ¿Cuáles son los límites de uso y para compartir?

Los límites de uso de los datos de Twitter incluyen restricciones en el uso comercial sin permiso, la prohibición de la redistribución masiva de datos y la necesidad de atribuir correctamente a los usuarios originales cuando se comparte su contenido. Estos límites son impuestos por Twitter para proteger la privacidad de los usuarios y garantizar un uso ético de la plataforma.

- ¿El usuario dispone de licencia?

Los usuarios de Twitter otorgan una licencia implícita para que sus mensajes sean accesibles públicamente en la plataforma. Sin embargo, no todos los usuarios están dispuestos a permitir la recopilación de sus datos, por lo que es importante respetar las preferencias de privacidad de cada usuario.

- ¿Qué tipo de uso está permitido?

Twitter permite el uso de datos con fines de investigación, lo cual se alinea con la categoría de «investigación» en el esquema de Cieri y DiPersio. Los investigadores pueden recopilar y analizar datos de Twitter para estudios académicos y científicos, siempre y cuando respeten las restricciones y políticas de la plataforma. Para ello, nos dimos de alta como investigadores académicos en la API de Twitter.

- ¿Qué tipo de procesamiento está permitido?

Los datos de Twitter pueden ser procesados para la creación de trabajos derivativos, como transcripciones de tuits o análisis de sentimientos. También es posible llevar a cabo un procesamiento transformador, como la construcción de modelos de lenguaje basados en los datos recopilados, lo que facilita la investigación en lingüística computacional y análisis de texto.

1.2.1. Selección de las cuentas de Twitter

Una vez establecida la necesidad de crear nuestro propio corpus, comenzamos por seleccionar las cuentas de Twitter a partir de las cuales crear nuestro corpus monolingüe, se realiza una pequeña investigación de cómo escriben los usuarios en gallego dentro de la plataforma. Las principales conclusiones extraídas son que hay bastante mezcla de ortografías usadas y distinto nivel de corrección de la lengua. Uno de los casos más llamativos son usuarios que optan por una ortografía reintregada, es decir, una ortografía más próxima a las reglas ortográficas del portugués que a la norma de la Real Academia Galega. Por esta razón, para garantizar un uso no solo correcto de la lengua, sino también idiomático y natural, se opta por seleccionar cuentas oficiales de las principales instituciones gallegas con actividad regular. En la tabla de la página siguiente, podemos ver las cuentas seleccionadas y el número de tuits descargados.

Cuenta de Twitter	Número de tuits
@uvigo	11 507
@UDC_gal	10 258
@UniversidadeUSC	7875
@AcademiaGalega	6362
@PortalPalabras	5543
@SXPL	4165

Tabla 3. Cuentas de Twitter junto con el número de tuits descargados para la creación del corpus monolingüe de redes sociales en gallego

1.2.2. *Script* de descarga

Para extraer el texto de los tuits y de los *hashtags*, utilizamos la librería de Python *snscrape*. Esta librería permite examinar los servicios de redes sociales (SNS) y descargar elementos como perfiles de usuarios o *hashtags*. También permite realizar búsquedas y descargar el resultado de estas, por ejemplo, las publicaciones relevantes. Esta librería se adecúa a nuestras necesidades, ya que nos permite indicar las cuentas específicas de las que queremos descargar todos los tuits. *Snscrape* devuelve un archivo en formato JSON, que convertimos también mediante un *script* de Python en un archivo CSV con el objetivo de poder manejar el contenido textual. El *script* permite descargar de manera individual cada una de las cuentas (ver Ilustración 30), por lo que se ejecuta tantas veces como cuentas escogidas para la creación del corpus.

Snscrape.py

```
import os
import pandas as pd
tweet_count = 100000
username = "Par_Gal"
os.system("snscrape --jsonl --max-results 100000 twitter-search
'from:Par_Gal'> text19.json".format(tweet_count, username))

# Reads the json generated from the CLI command above and creates a pandas
dataframe
tweets_df1 = pd.read_json('text19.json', lines=True)
tweets_df1.to_csv('text19.csv', sep=',', index=False)
```

Ilustración 30. Script utilizado para descargar los tuits de las cuentas escogidas en un archivo CSV

Una vez tenemos creado el archivo CSV (ver Ilustración 31, página 102), eliminamos todo el contenido menos el texto de los tuits y de los *hashtags*. Posteriormente, se efectúa

una limpieza para eliminar URL e iconos. Finalmente, se ejecuta una búsqueda manual de contenido en otras lenguas.

1.2.3. La traducción inversa

Tras crear el corpus monolingüe, el siguiente paso consiste en traducirlo al castellano para obtener un corpus bilingüe. Para ello, utilizamos el traductor neuronal genérico de *Google Translate* (Wu *et al.*, 2016) en la herramienta SDL Trados Studio[17]. El objetivo es generar de una manera automática el castellano, ya que lo que queremos priorizar es la calidad del idioma de llegada, en este caso, el gallego. Una vez traducido, se exporta el contenido a una memoria de traducción en formato TMX (Translation Memory eXchange) que es el formato estándar para el intercambio de memorias de traducción. Es un formato admitido por la práctica totalidad de las herramientas de traducción asistida. A partir de la memoria creada por nosotros y el corpus *Paracrawl* se empieza a procesar el corpus con el formato necesario para entrenar un motor.

Ilustración 31. CSV exportado directamente con nuestro *script* sin limpiar

1.2.4. Procesamiento del corpus

Los motores de traducción automática necesitan los corpus en un formato específico. Habitualmente, parten de un archivo de texto separado por tabulador en el que estén alineados los dos idiomas. Posteriormente, es necesario limpiar este archivo de texto y *tokenizarlo* para que el motor lo pueda procesar. Todos estos pasos se realizan con otra librería de Python llamada MTUOC. El proyecto MTUOC (Oliver, 2020) distribuye una

[17] Disponible en: https://www.trados.com/. Fecha de consulta: 15 de agosto de 2023

serie de componentes que permiten entrenar y utilizar sistemas de TAN y estadística en ordenadores de consumo, independientemente de si utilizan un sistema operativo Linux, Windows o Mac. El proyecto cuenta con una serie de programas y módulos de Python que permiten la *tokenización*, *truecasing* y limpieza del corpus, así como todos los pasos de preprocesamiento de corpus con un solo comando.

Una vez tenemos las memorias convertidas en el formato de texto de trabajo, ejecutamos el *script* de limpieza de corpus del proyecto MTUOC que permite asegurarse de eliminar los siguientes elementos:

- Presencia de caracteres representados por su entidad html/xml.
- Segmentos vacíos en alguna de las lenguas.
- Segmentos en lenguas diferentes de las requeridas.
- Segmentos con un porcentaje alto de expresiones numéricas.
- Segmentos con etiquetas XML.

Tras la limpieza del corpus, el siguiente paso es combinar el corpus específico con el corpus genérico. Para ello, se emplea un componente del proyecto MTUOC que tiene como finalidad combinar dos corpus paralelos en uno solo. Este módulo es especialmente útil cuando el corpus paralelo de especialidad tiene un tamaño insuficiente para entrenar un sistema de traducción automática y disponemos de un corpus paralelo de tamaño mucho mayor para el mismo par de lenguas que nos puede permitir completar el corpus paralelo de especialidad. El programa selecciona un número determinado de segmentos del corpus de gran tamaño que sean lo más parecidos posible a los del corpus de especialidad. Para seleccionar los segmentos más parecidos el programa calcula un modelo de lengua a partir de los segmentos en la lengua de partida y selecciona los segmentos del corpus de gran tamaño cuya perplejidad de los segmentos en la lengua de partida sean menores.

Finalmente, utilizamos otro *script* que nos permite ejecutar los pasos básicos de preprocesamiento de corpus para un entrenamiento neuronal descritos en el marco teórico: *tokenización*, *truecasing*, tratamiento de expresiones numéricas, tratamiento de emails y URL, subpalabras y división del corpus en *train*, *eval* y *val*. En nuestro caso, realizamos todos los pasos y procesamos el archivo con *sentencepiece*. En el Anexo II, se pueden ver los elementos que conforman el archivo del *script*.

Una vez ejecutado el *script*, estos son los archivos necesarios para iniciar un entrenamiento:

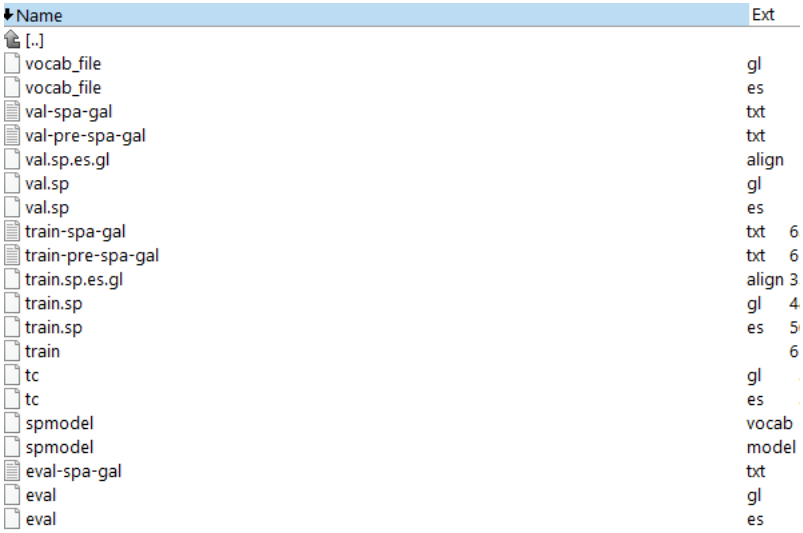


Ilustración 32. Ejemplo de listado de archivos necesarios para iniciar un entrenamiento neuronal

Los archivos *train* contienen el conjunto de datos que se utilizan para entrenar el modelo. Los datos en este conjunto se utilizan para ajustar los parámetros del modelo a medida que el modelo aprende de ellos. El conjunto de validación, archivo *val*, se utiliza para ajustar hiperparámetros, como la tasa de aprendizaje o la arquitectura del modelo, con el objetivo de mejorar el rendimiento del modelo en datos no vistos. Se realiza una validación cruzada iterativa entre el conjunto de validación y el conjunto de entrenamiento para afinar el modelo. Por último, el archivo de evaluación, *eval*, contiene un conjunto de datos que se utiliza durante el proceso de entrenamiento para evaluar el rendimiento del modelo en datos que no ha visto previamente.

1.3. Consideraciones finales

En este capítulo, hemos detallado el proceso integral de creación de nuestro propio corpus de tuits a partir de cuentas institucionales gallegas. A lo largo de este proceso, hemos enfrentado desafíos técnicos y metodológicos que requerían soluciones creativas y rigurosas.

Uno de los aspectos más destacados de esta creación de corpus fue la selección estratégica de las cuentas institucionales. La elección de estas cuentas se basó en criterios específicos que garantizaban el buen uso de la lengua gallega, lo que aporta riqueza y relevancia a

nuestro corpus. Además, la captura de datos se realizó de manera sistemática, siguiendo las pautas éticas y legales correspondientes. Posteriormente, se utilizó la técnica de traducción inversa para conseguir un corpus bilingüe castellano-gallego. Finalmente, unimos los dos corpus disponibles para iniciar el entrenamiento del motor.

En resumen, el proceso de creación de nuestro corpus a partir de tuits de cuentas institucionales gallegas es un componente fundamental de nuestra investigación. Este recurso nos proporciona una base sólida para llevar a cabo el entrenamiento del motor. En los siguientes capítulos, nos centraremos en los análisis y resultados derivados de este corpus.

CAPÍTULO VII: Metodología de entrenamiento y evaluación del motor: la prueba piloto

Tras la irrupción de los modelos de lengua, la traducción automática neuronal (TAN) ha mejorado, sobre todo, en términos de fluidez. Estos sistemas neuronales han superado muchas de las dificultades que aquejaban a los sistemas de traducción automática anteriores basados en reglas y estadísticos (Hassan *et al*., 2018), lo que ha llevado a afirmaciones de que la TAN alcanza la paridad con las traducciones humanas (Toral et *al*., 2018). Como hemos visto en el marco teórico, la forma más común de determinar esta paridad es mediante el uso de métricas automáticas de evaluación de calidad que comparan el resultado directo de la traducción automática con la traducción humana del mismo texto de origen (Chatzikoumi, 2020). Sin embargo, estas diferencias entre las dos traducciones pueden deberse a dos problemas: o bien una de las traducciones contiene errores, o bien las dos traducciones son completamente correctas a pesar de las diferencias en la formulación.

Por esta razón, en esta investigación tenemos el objetivo de acercarnos a la evaluación de la traducción automática desde la premisa de que las traducciones de un mismo texto pueden diferir y, por lo tanto, nuestra hipótesis es que la calidad de la traducción automática debe evaluarse también utilizando un principio de comparación diferente. Comprobaremos nuestra hipótesis en los siguientes capítulos.

1. Elección de la tecnología TAN

En la fase inicial de investigación, se opta por una tecnología con fines educativos con la que entrenar una primera versión del motor a modo de piloto para detectar posibles problemas. Posteriormente, tal y como se explica en el capítulo VI buscamos otra solución más robusta y profesional pensada para el entrenamiento definitivo del motor. La tecnología neuronal escogida para el primer entrenamiento del motor fue JoeyNMT[18]. Se trata de una herramienta para TAN que pretende ser «*minimalist NMT for educational purposes*». El objetivo es crear una herramienta que implemente las principales técnicas

[18] Disponible en: https://github.com/joeynmt/joeynmt. Fecha de consulta: 15 de agosto de 2023.

de TAN con un código sencillo que permita modificarlo fácilmente. JoeyNMT fue desarrollado inicialmente y actualizado por Jasmijn Bastings (University of Amsterdam) y Julia Kreutzer (Heidelberg University), ambas investigadoras ahora en Google Research. Actualmente, Mayumi Ohta de la Heidelberg University es quien continúa con el testigo. JoeyNMT se ejecuta a través de la plataforma MutNMT[19] desarrollada dentro del proyecto MultiTraiNMT - *Machine Translation training for multilingual citizens* (2019-1-ES01-KA203-064245, 01/09/2019–31/08/2022). El objetivo de este proyecto es desarrollar, evaluar y divulgar materiales de libre acceso que permitan la mejora de la enseñanza y aprendizaje sobre traducción automática. MutNMT proporciona las siguientes funcionalidades:

- Subir y gestionar corpus
    - Subir corpus en formato de texto, TMX o TSV
    - Categorizar el corpus en función del dominio
    - Compartir el corpus con otros usuarios
- Entrenar y gestionar motores
    - Seleccionar corpus o un subconjunto de ese corpus para entrenar un motor con un modelo *transformer*
    - Monitorizar el proceso de entrenamiento con tablas de datos y gráficos
    - Parar, retomar y reiniciar el proceso de entrenamiento en cualquier momento
- Traducir texto y documentos
- Evaluar las traducciones con las métricas BLEU, chrF3, TER y TTR

Se opta por esta plataforma dado que permite entrenar el motor empleando una interfaz gráfica, lo que facilita el proceso de entrenamiento al evitar ejecutar el entrenamiento por pantalla de comandos. También se tiene que en cuenta que permite utilizar los servidores GPU de la UAB para el entrenamiento y no requiere buscar servidores propios donde instalar la tecnología JoeyNMT.

[19] Disponible en: https://ntradumatica.uab.cat. Fecha de consulta: 15 de agosto de 2023

2. Proceso de entrenamiento

El primer paso para poder entrenar el motor de la prueba piloto en la plataforma MutNMT es subir el corpus de entrenamiento usando una cuenta de experto. La plataforma permite subir el corpus preprocesado en formato .tmx o .csv, o bien subir en formato .txt dos corpus monolingües previamente alineados. La pestaña de subida de corpus simplemente te solicita que introduzcas un título, una descripción, un dominio, las lenguas del corpus y el archivo, tal y como se ve en la Ilustración 33. En nuestro caso, subimos el corpus preprocesado previamente con los módulos MTUOC.

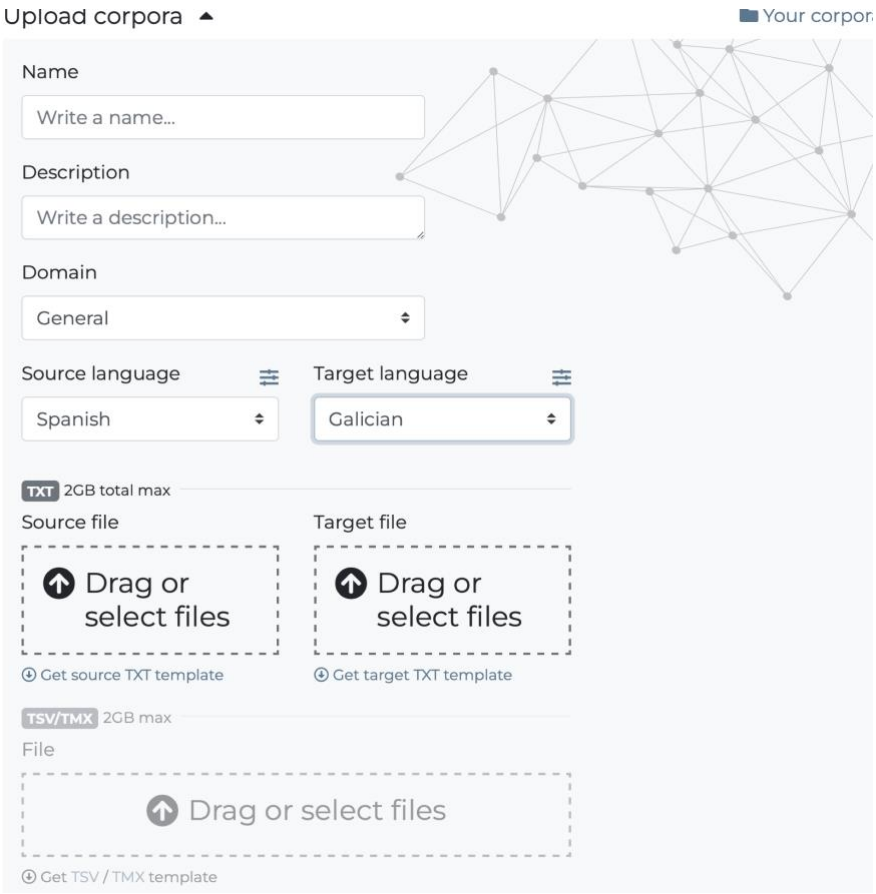


Ilustración 33. Pantalla de subida de corpus en la plataforma MutNMT

Una vez el corpus está dentro de la plataforma, el siguiente paso es definir los parámetros de entrenamiento. Tal y como vemos en la Ilustración 34, en esta pestaña se debe introducir el nombre del motor, elegir los idiomas, escribir una descripción si lo consideramos pertinente, definir el número de *epochs (*un paso completo de entrenamiento de un conjunto de oraciones del corpus de entrenamiento, en esta plataforma se recomienda escoger entre 7 y 10 *epochs*), las condiciones de parada (el

motor se parará si no mejora después del número indicado de validaciones), la frecuencia de validación (la cantidad de pasos incluidos antes de una evaluación del estado del entrenamiento y que debe estar entre 3000 y 9000), el tamaño del vocabulario (la cantidad de palabras del vocabulario debe estar entre 16 000 y 32 000), el tamaño del lote (la cantidad de *tokens* procesados en cada paso y que debe estar entre 6000 y 9000 *tokens*), el número de hipótesis de traducción para una palabra y el tamaño de cada uno de los corpus de entrenamiento, validación y evaluación.

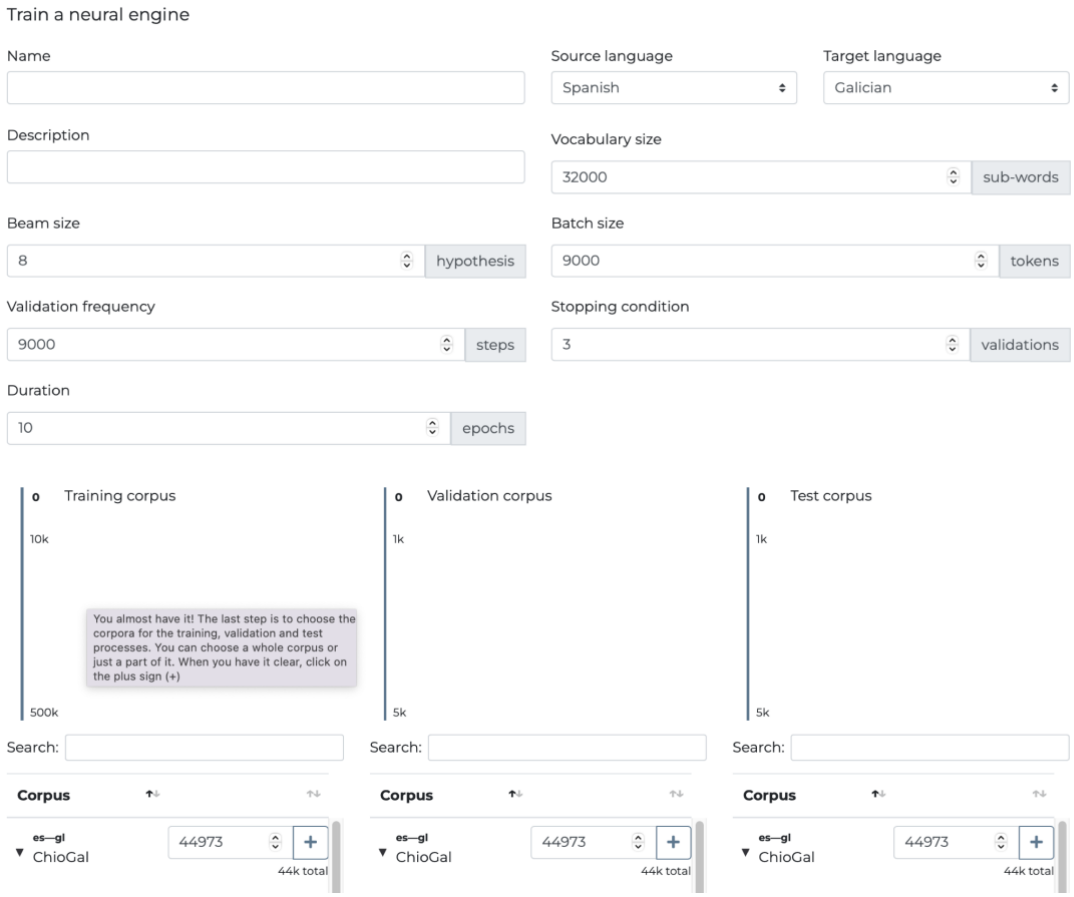


Ilustración 34. Parámetros de entrenamiento de la plataforma MutNMT

Una vez definidos los parámetros, se inicia el proceso de entrenamiento. Por defecto, la plataforma para cada hora completada, por lo que se tiene que reiniciar el proceso hasta conseguir un óptimo resultado. En nuestro caso, entrenamos el motor 11 horas siguiendo las recomendaciones de parámetros de la plataforma y obtuvimos un BLEU de 70.47 (ver Ilustración 35). No se sigue entrenando más horas porque el motor no mejora.

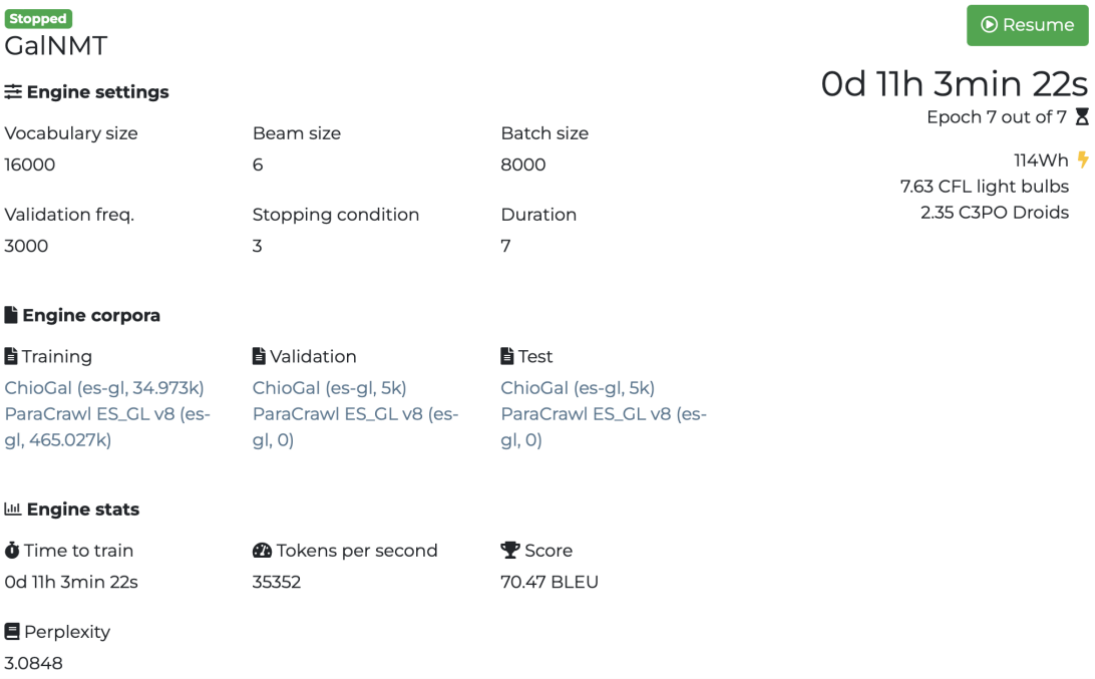


Ilustración 35. Parámetros de entrenamiento de nuestro motor en la plataforma MutNMT

Tras estas horas de entrenamiento, ya se puede traducir con nuestro motor desde la pestaña de traducción de la plataforma MutNMT (ver Ilustración 36).

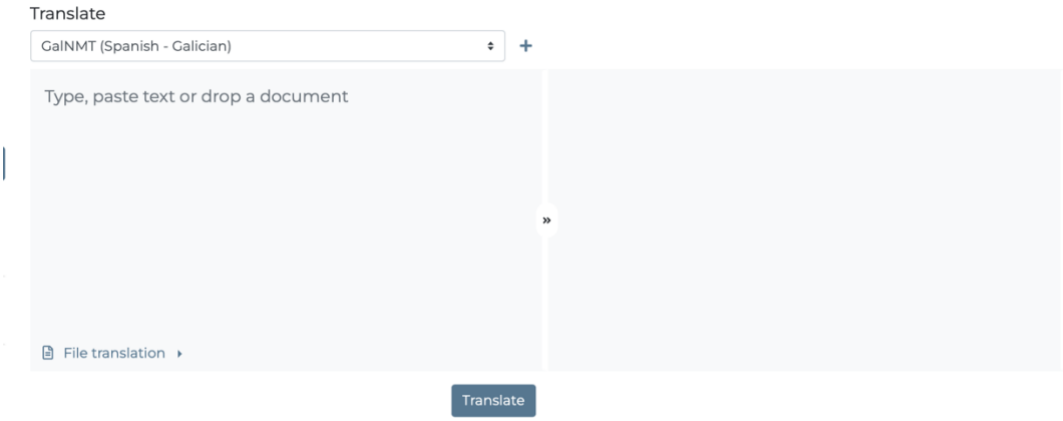


Ilustración 36. Pestaña para traducir con el motor creado en la plataforma MutNMT

Una vez tenemos el motor entrenado, el siguiente paso es evaluarlo. Para ello, evaluaremos de manera automática el motor (véase apartado siguiente) y realizaremos una prueba piloto con el doble objetivo de evaluar el motor y validar nuestra propuesta de evaluación (véase capítulo VIII).

3. Evaluación automática

Para efectuar la evaluación automática del motor, empleamos BLEU (Bilingual Evaluation Understudy), ya que es una de las métricas automáticas más ampliamente utilizadas y estudiadas y es la métrica que la plataforma MutNMT da como defecto. Esta métrica busca cuantificar la calidad de una traducción comparándola con una o más traducciones de referencia humanas. BLEU se originó en el trabajo pionero de Papineni *et al.* (2002), quienes propusieron una medida de coincidencia de n-gramas entre la traducción candidata y las referencias humanas. A lo largo de los años, el BLEU ha sido objeto de numerosos estudios y mejoras, y ha sido ampliamente adoptado como una herramienta estándar para la evaluación automática en la comunidad de NLP (*Natural Language Processing*, por sus siglas en inglés). A pesar de su popularidad y éxito relativo, BLEU no está exento de críticas y limitaciones, como su enfoque en la coincidencia de n-gramas y su insensibilidad a la coherencia y la fluidez del texto, por lo que completaremos esta información con la evaluación cualitativa que describiremos en el siguiente capítulo.

Antes de visualizar los resultados obtenidos, cabe destacar que el motor se evalúa utilizando un corpus de validación extraído exclusivamente del corpus específico de Twitter para asegurarnos de evaluar la adecuación al dominio de investigación. Este corpus de validación no se usó para entrenar el motor. Asimismo, para poder entender la métrica, debemos aclarar que se expresa habitualmente en un porcentaje que va del 0 al 100 %, siendo el 0 lo más alejado de la traducción de referencia y 100 % la exactitud total. Cuanto mayor sea el porcentaje de BLEU, se entiende una mayor calidad de la traducción.

Tipo de motor	BLEU
JoeyNMT	70,63 %

Tabla 4. Evaluación automática BLEU

Si comparamos este valor con motores similares de la bibliografía consultada, constatamos que el porcentaje BLEU obtenido es superior a la media que, por ejemplo, en el caso de Ortega *et al.* (2022: 93), se sitúa en torno al 50 %, lo que de entrada parece indicar un buen rendimiento ya en la primera versión del motor.

4. Consideraciones finales

Tal y como avanzamos en el marco teórico de esta investigación, para poder entender el rendimiento de un motor es necesario combinar evaluaciones automáticas con evaluaciones humanas. Dentro de este capítulo, hemos avanzado la estrategia de evaluación que seguiremos y validaremos en la prueba piloto y hemos descrito el resultado obtenido en el primer motor entrenado en la evaluación automática mediante la métrica BLEU. Como se ha podido constatar, el uso de un corpus específico más grande y de un mayor control en la configuración del motor, mejora el rendimiento del motor y arroja un porcentaje BLEU, ya de por sí elevado, si lo comparamos con motores a lenguas similares de la bibliografía consultada.

CAPÍTULO VIII: La evaluación basada en no inferioridad: resultados de la prueba piloto

En el capítulo anterior, se describe la primera versión del motor entrenado y se avanzan los resultados obtenidos en la evaluación automática. El enfoque particular de nuestra investigación es proponer la evaluación de la traducción automática utilizando el principio estadístico de la no inferioridad. Dado que se trata de un enfoque innovador, se lleva a cabo primero una prueba piloto con la primera versión del motor para poder evaluar nuestro instrumento de análisis y, posteriormente, realizar un estudio para verificar la no inferioridad. Por esta razón, este capítulo lo dedicaremos a describir la prueba piloto. Posteriormente y tras analizar las conclusiones de esta prueba piloto, en los capítulos IX y X, describiremos el nuevo entrenamiento realizado y la nueva evaluación tanto cuantitativa como cualitativa, esta última ya desde dos perspectivas: la evaluación humana de la traducción automática realizada por lingüistas profesionales (apartado 2 del capítulo X) y la evaluación de la recepción de las traducciones automáticas por partes de hablantes de la lengua gallega (apartado 3 del capítulo X). La evaluación humana se describirá en el capítulo X tras el entrenamiento final del motor,

La evaluación por parte de los hablantes se basa en las respuestas obtenidas en un cuestionario y se analizan a partir del principio de no inferioridad. A través de este principio, pretendemos determinar si los textos obtenidos a través de la traducción automática son menos naturales que los textos creados en ese mismo idioma y extraer percepciones del usuario final. La idea es utilizar este principio para entender mejor la evaluación automática BLEU y la evaluación humana de clasificación de errores (marco MQM-DQF) efectuadas para así tener una visión completa de la calidad del motor, tal y como veremos en el análisis de resultados de los capítulos posteriores.

Primero se recopilan muestras representativas de tuits escritos originalmente en gallego y tuits escritos en castellano que se tradujeron empleando el motor de traducción automática desarrollado. Como se explicará más en detalle a continuación, los tuits se clasifican según su longitud. Posteriormente, se diseñan distintos cuestionarios y se distribuyen entre usuarios de Twitter gallegos para extraer su percepción. Finalmente, se

analizan los datos con la ayuda del servicio de estadística de la Universitat Autònoma de Barcelona.

Los resultados obtenidos durante el piloto proporcionan una visión detallada sobre la calidad de las traducciones generadas por el motor y la solidez del método de evaluación, lo que nos permite identificar posibles áreas de mejora. Así, la información recopilada durante el piloto es de gran relevancia para el proceso de desarrollo y optimización del motor de traducción automática. Los hallazgos y las conclusiones extraídas a partir de este piloto proporcionan una base sólida para realizar mejoras y ajustes necesarios en el motor, con el objetivo de lograr traducciones automáticas de mayor calidad y precisión.

1. El principio estadístico de no inferioridad

El principio estadístico de no inferioridad ha demostrado ser una herramienta valiosa en diversos campos, como la farmacología, donde se utiliza para evaluar la equivalencia de medicamentos genéricos en comparación con medicamentos patentados (Molina, 2020; Althunian *et al*., 2017, Tunes da Silva, 2009). Sin embargo, su aplicación en la evaluación de la traducción automática es un enfoque relativamente novedoso y prometedor.

El principio de no inferioridad se basa en la hipótesis nula de que la diferencia entre dos tratamientos o intervenciones es menor o igual a un margen predefinido de no inferioridad. Es decir, si la intervención «B» se considera no inferior a la intervención «A», se espera que la diferencia en los resultados entre ambas intervenciones sea menor o igual a un umbral aceptable de no inferioridad. Molina concluye que (2020:55):

> *Es necesario definir a priori el margen de no inferioridad. La selección de este margen es clave y se basa en una combinación entre el criterio clínico y el análisis estadístico de los datos de eficacia del tratamiento que se utiliza como comparador respecto a placebo.*
>
> *Cuánto más amplio es el margen de no inferioridad (valor delta: δ o Δ), más se reduce el tamaño muestral del ensayo y más fácil es concluir la no inferioridad.*

Por su parte, Althunian *et al*. (2017: 1636) afirman que la recomendación habitual es «*to compare the estimated 95% confidence interval (CI) of the new drug vs. the active comparator from the noninferiority trial to a predefined margin*».

La aplicación del principio de no inferioridad en la evaluación de la traducción automática implica establecer un margen de no inferioridad, que representa la diferencia aceptable

entre las traducciones generadas por el sistema automático y las traducciones humanas. Si la diferencia se encuentra dentro de este margen, se concluye que el sistema automático produce traducciones no inferiores a las traducciones humanas. Como veremos en la prueba piloto y en la evaluación final, definiremos márgenes de no inferioridad exigentes, en línea con lo recomendado en farmacología. Este enfoque de evaluación basado en el principio de no inferioridad tiene varias ventajas. En primer lugar, permite establecer criterios claros y objetivos para evaluar la naturalidad y aceptación de las traducciones automáticas. En segundo lugar, evita la necesidad de alcanzar la perfección absoluta en la traducción automática, ya que se centra en garantizar que las traducciones no sean significativamente inferiores a las traducciones humanas.

El análisis presentado en este capítulo se basa en unos cuestionarios realizados a usuarios gallegos de Twitter, a quienes se les pidió evaluar su percepción sobre una muestra aleatoria de tuits traducidos del castellano al gallego, así como tuits escritos directamente en gallego. Se utilizó la primera versión del motor para obtener las traducciones de los tuits. Nuestro objetivo, por tanto, no es solo evaluar el motor de traducción automática desarrollado para esta prueba piloto, sino también demostrar la validez del enfoque de no inferioridad empleado para evaluar la calidad de la traducción automática. De conseguir validar esto, el siguiente objetivo es utilizar la percepción de los usuarios finales para encontrar un vínculo entre los errores detectados en la evaluación humana siguiendo el marco MQM-DQF y una actitud negativa hacían los tuits traducidos automáticamente, de manera similar a la metodología aplicada por Guerberof *et al.*, 2022 y Bhardwaj *et al.*, 2020.

2. Diseño del estudio

2.1. Descripción de la muestra

Para diseñar el estudio de la prueba piloto basado en el principio de no inferioridad, optamos por un estudio empírico con un diseño factorial incompleto. Configuramos cuatro cuestionarios usando el programa Google Forms con un total de 20 ítems (tuits) de una muestra posible de 60 tuits. Esta muestra está compuesta por 30 tuits escritos originalmente en gallego extraídos de las mismas cuentas usadas para la creación del corpus, pero que no se emplearon en los datos de entrenamiento y 30 tuits de la misma temática escritos originalmente en castellano y traducidos automáticamente con nuestro

motor. Además, también dividimos cada una de las categorías anteriores según su longitud. Se decide introducir esta diferenciación para poder analizar el rendimiento del motor según la longitud del segmento, dado que uno de los principales desafíos para la TAN es la correcta resolución de frases muy cortas o largas; las primeras debido a la falta de contexto y las segundas debido a su complejidad lingüística. Para ello, tomamos como referencia Castilho *et al.* (2017) y Sánchez-Gijón *et al.* (2019), pero adaptándolo a nuestro dominio para tener en cuenta las restricciones de caracteres inherentes a Twitter.

De esta forma, podemos distinguir:

- Tuits formados por una frase corta delimitada con un mínimo de dos palabras y un máximo de diez palabras.
- Tuits formados por una frase larga de más de diez palabras.
- Tuits de un párrafo compuesto por dos o más frases cortas de un mínimo de dos palabras y un máximo de diez de palabras.
- Tuits de un párrafo formado por dos o más frases largas de más de diez palabras.
- Tuits de párrafo mixto en el que podíamos encontrar un mínimo de una frase corta de hasta diez palabras y un mínimo de una frase larga de más de diez palabras.

2.2. Composición de los cuestionarios

Para elaborar los cuestionarios, formamos dos bloques: uno de tuits originales y otro de tuits traducidos, compuestos a su vez por seis tuits de cada una de las longitudes anteriores. De manera deliberada, se busca que la proporción de aparición de los tuits originales y traducidos no fuera igual para evitar que el participante pudiese deducirlo por el simple formato del cuestionario. Así mismo, para garantizar la validez de las respuestas, cada cuestionario comparte 10 ítems con los otros, cinco con un cuestionario y otros cinco con otro cuestionario. En el Anexo III, se puede consultar la distribución de cada uno de los cuestionarios.

2.3. Formato de los cuestionarios

Para la evaluación de los datos del motor creado con JoeyNMT, preparamos cuatro cuestionarios formados por tres secciones. En la primera sección, se ofrece al participante una explicación con las instrucciones necesarias para realizar el cuestionario y el propósito de esta. En la siguiente sección, se recogen algunos datos demoscópicos de los participantes como son: sexo, edad, lectura y escritura habitual en lengua gallega y uso

habitual de traducción automática. En la última sección dedicada específicamente a los tuits, se muestran los tuits de uno a uno para evitar que los participantes anticipen algún patrón sistemático. Los tuits se acompañan de la siguiente pregunta de respuesta sí/no:

¿Consideras que este chío está redactado de forma orixinal en galego ou que foi traducido por un motor de tradución automática?

Ilustración 37. Pregunta de respuesta sí/no del cuestionario de la prueba piloto

Las opciones que los participantes tienen son cerradas e inequívocas:

Si, considero que está redactado orixinalmente en galego
Non, considero que é unha tradución automática ao galego

Ilustración 38. Opciones de respuesta a la pregunta principal del cuestionario de la prueba piloto

2.4. Perfil de los participantes

En la prueba piloto, participaron un total de 261 personas. Los participantes debían ser hablantes de lengua gallega y usuarios de Twitter. Los cuestionarios se distribuyeron a través de Twitter, dentro de comunidades especialmente activas en el uso y promoción del gallego. En la tabla de la página siguiente, podemos ver el resumen de los datos descriptivos de los participantes según sexo, edad, uso habitual del gallego y uso de la traducción automática en su día a día.

	[ALL]	N
	N=261	
Sexo:		261
Feminino	136 (52.1%)	
Masculino	116 (44.4%)	
Non binario	3 (1.15%)	
Prefiro non dicilo	6 (2.30%)	
Idade	39.9 (12.4)	256
Galego:		261
Non	53 (20.3%)	
Si	208 (79.7%)	
Emprega T.Automática:		261
Non	220 (84.3%)	
Si	41 (15.7%)	

Tabla 5. Resumen descriptivo de los datos demográficos

Como vemos, el perfil de participantes es un hablante, de aproximadamente 40 años, que conscientemente elige utilizar como lengua vehicular el gallego en su día a día y que no usa habitualmente la traducción automática.

3. Procesamiento de los datos

Nuestra aproximación estadística se basa en el principio de no inferioridad, es decir, si la TAN hubiese igualado la calidad de la traducción humana, se daría una percepción equiparable de la naturalidad entre un tuit escrito originalmente en gallego y un tuit traducido con nuestro motor. Esta percepción del usuario de Twitter se calcula mediante la tasa de aceptación como original o no de cada uno de los tuits.

El análisis estadístico se lleva a cabo mediante el software R v4.1. Para describir las variables cuantitativas, se usan los índices promedio, desviación estándar, máximo,

mínimo y número de casos. Por su parte, para describir variables cualitativas se emplean frecuencias absolutas y relativas.

La comparación preliminar entre la tasa de aceptación y las características de los tuits o las características demográficas de los participantes ha calcula empleando la prueba χ^2 o la prueba Exacta de Fisher según el cumplimiento de los criterios de Cochran.

Se fija el nivel de significación para todas las pruebas estadísticas en el 5 % ($p<0.05$).

Para comparar la tasa de aceptación de los tuits traducidos automáticamente se emplea un modelo lineal generalizado mixto con distribución binomial, contemplando el tuit (ítem) y el participante (Id) como factores aleatorios cruzados, así como las características de los tuits como factores fijos. Por tanto, se obtiene una discriminación de los datos según el participante o según el origen del tuit. En cada una de las aproximaciones a los resultados que veremos en los siguientes puntos, se indicará el detalle estadístico correspondiente. El modelo empleado es el siguiente:

$$\log\left[\frac{p}{1-p}\right] = \alpha + \beta X + a_i + b_j$$
$$a_i \sim N(0, \sigma^2_{ITEM})$$
$$b_j \sim N(0, \sigma^2_{ID})$$

Ecuación 1. Modelo estadístico empleado en la prueba piloto

siendo p la probabilidad de aceptación como original o no; X la matriz de diseño de los efectos fijos, es decir, la longitud y origen del tuit; β el correspondiente vector de coeficientes y a_i, b_j los efectos aleatorios para Ítem (tuit) e Id (participante).

4. Análisis de los datos

El principal objetivo de la evaluación de los resultados de la prueba piloto a través del principio de no inferioridad es validar el método, la fórmula y los instrumentos de recogida de datos. Buscamos, por tanto, en primer lugar, confirmar si la no inferioridad se puede aplicar en este contexto y también, si la cantidad de datos y recogida de los mismos son suficientes. En segundo lugar, también pretendemos determinar la propia

calidad del motor. Para ello, primero analizamos la no inferioridad global y, posteriormente, según el tipo de tuit, teniendo en cuenta no solamente los datos estadísticos (cálculo de la no inferioridad) sino también la distribución (evaluación de la no inferioridad según respuesta).

4.1. Resultados globales de la prueba piloto

El tamaño de muestra inicial del estudio es de 261 participantes, totalizando 5220 valoraciones. Para poder entender las percepciones de los participantes con respecto a la traducción automática de nuestro motor, debemos primero establecer que la media de tuits originales percibidos correctamente como tales se sitúa en el 68,7 % (ver Ilustración 39). Esto quiere decir que algunos tuits originales se percibieron como traducción automática, por lo que hay que tener en cuenta este dato a la hora de calcular la no inferioridad. Podríamos entonces considerar la TAN como no inferior a los tuits originales si alcanza un porcentaje similar de aceptación como original que se sitúe en torno a ese casi 70 %. Así, creamos el siguiente gráfico de dispersión para poder ver dónde se sitúa la tasa de aciertos de la TAN y donde la del texto original.

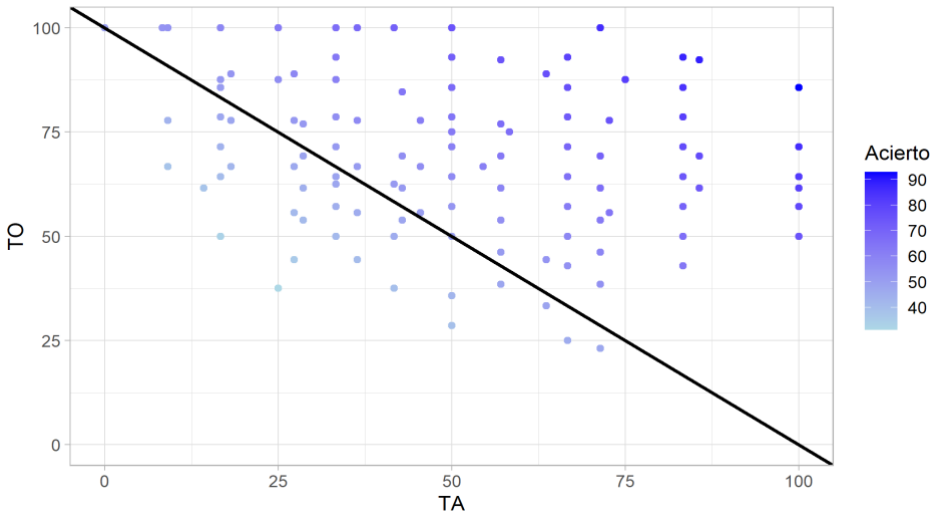


Ilustración 39. Tasa de aciertos en TA vs Tasa de aciertos en TO

La línea negra marca las valoraciones que no discriminan, esto es, en los puntos de la línea la tasa de aciertos es del 50 % y es equivalente a valoraciones realizadas al azar. A su vez, los puntos por encima de la línea negra corresponden a sujetos que sí discriminan, es decir, que pueden distinguir entre TA y TO y el acierto global superior al 50 %. Por el contrario, los puntos por debajo de la línea negra corresponden a sujetos que discriminan peor que el azar, es decir, valoran erróneamente. En cierta medida tienen tendencia a

clasificar los TA como TO; y los TO los clasifican como TA. Nótese que, si la herramienta de traducción automática fuera perfecta, los sujetos no podrían discriminar y sus tasas estarían cerca del 50 %, con valoraciones en la línea negra, por lo que, a nivel global, el motor de TAN no es lo suficientemente bueno para alguno de los participantes.

4.2. Resultados por tipo de tuit

Tal y como describimos en la sección 2.1, clasificamos los tuits en función de su longitud. De esta forma, podemos diferenciar: tuits formados por una frase corta, tuits formados por una frase larga, tuits formados por un párrafo de frases cortas, tuits formados por un párrafo de frases largas y tuits formados por un párrafo de frases largas y cortas. Si analizamos la aceptación según el tipo de tuit, observamos que hay una tendencia bastante clara en rechazar aquellos tuits compuestos por frases cortas. En este tipo de tuits, la distancia entre traducción automática y original es mayor que en las otras variables. Esta debilidad en las frases cortas es algo común y frecuente en la TAN y habitualmente se suele achacar a la falta de contexto. No obstante, comprobamos que estos tuits son gramaticalmente correctos, por lo que entendemos que no hay suficiente información para que el participante pueda discriminar si suena natural o no en gallego, especialmente dada su cercanía con el español.

Por contra, observamos que los tuits traducidos automáticamente que se componen de párrafos con frases largas tienen una tasa de aceptación mayor que los originales de su misma categoría. De manera similar, se puede apreciar la misma tendencia en los párrafos mixtos. Una de las posibles razones de esta mayor aceptación podría ser que la no restricción de caracteres en la traducción automática y una mayor fluidez.

Para analizar la homogeneidad en las respuestas según las subcategorías que previmos, creamos el siguiente diagrama de caja y bigotes.

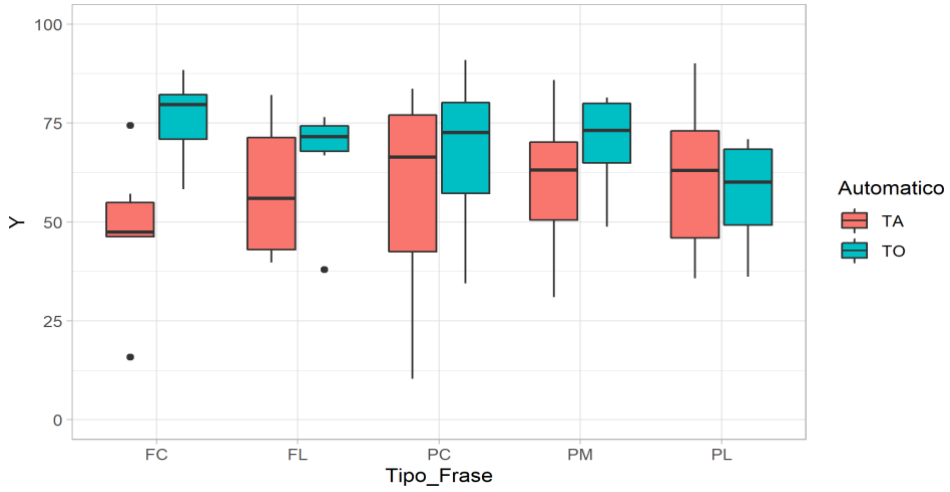


Ilustración 40. Diagrama de caja y bigotes según la aceptación del tuit

Lo que podemos apreciar en esta ilustración es que se da una mayor dispersión en la traducción automática que en los tuits redactados originalmente. Como podemos observar, las cajas en los textos originales son más pequeñas que en la traducción automática, incluso en los tuits de párrafos de frases largas que han sido más aceptados. Esto nos lleva a concluir que se da un mayor acuerdo en cómo son los tuits escritos originalmente que en la traducción automática.

4.3. Análisis estadístico

Para realizar el análisis estadístico, en primer lugar, se modeliza, es decir, se desarrolla un modelo estadístico que describe la relación entre la aceptación de los tuits y las características de estos a través de un modelo lineal generalizado mixto con función de enlace *logit*. La función de enlace *logit* es una función matemática que se utiliza en el contexto de la regresión logística para modelar y predecir la probabilidad de un evento binario (sí/no, 0/1) en función de un conjunto de variables predictoras. La función de enlace logit transforma la escala de probabilidad lineal (que puede variar en un rango infinito) en una escala logarítmica, lo que permite modelar de manera efectiva la relación entre las variables predictoras y la probabilidad de que ocurra el evento de interés. Esta propuesta modeliza el logaritmo del *Odds* de aceptación en función de diferentes variables. El *Odds* de aceptación viene a ser una transformación de las probabilidades, de forma que, si el *Odds* es 0, las probabilidades son del 50 %, y si el *Odds* es positivo, las probabilidades son mayores del 50 %. Al comparar grupos se obtiene el *Odds* Ratio,

medida similar a una tasa de riesgos (probabilidades). Si el *Odds* Ratio es 1, no hay diferencias entre grupos. Si es mayor que 1, el primer grupo tiene una incidencia mayor.

Tal y como vemos en la Ilustración 41, primero se crea el modelo (línea *generalized linear model*) y se determinan los efectos aleatorios (línea *random effects*) y los efectos fijos (línea *fixed effects*). Por último, se calcula la devianza (tabla *Analysis of Devianza*) para evaluar la significancia de las primeras estimaciones del modelo.

```
m0 <- glmer( Aceptación  ~ Automatico + Tipo_Frase +  Automatico : Tipo_Frase
+ (1|ID) +  + (1|Item) , family=binomial(link='logit'), data=bd3)
print( summary(m0),  correlation = FALSE)

  Generalized linear mixed model fit by maximum likelihood (Laplace
    Approximation) [glmerMod]
   Family: binomial  ( logit )
  Formula: Aceptación ~ Automatico + Tipo_Frase + Automatico:Tipo_Frase +
      (1 | ID) + +(1 | Item)
     Data: bd3

       AIC      BIC   logLik deviance df.resid
    6239.8   6318.5  -3107.9   6215.8     5208

  Scaled residuals:
      Min      1Q  Median      3Q     Max
  -3.2296 -0.8603  0.4642  0.6728  2.9437

  Random effects:
   Groups Name        Variance Std.Dev.
   ID     (Intercept) 0.2520   0.5020
   Item   (Intercept) 0.5878   0.7667
  Number of obs: 5220, groups:  ID, 261; Item, 60

  Fixed effects:
                             Estimate Std. Error z value Pr(>|z|)
  (Intercept)                 -0.1559     0.3350  -0.465  0.64164
  AutomaticoTO                 1.4060     0.4750   2.960  0.00307 **
  Tipo_FraseFL                 0.4806     0.4752   1.011  0.31187
  Tipo_FrasePC                 0.3833     0.4804   0.798  0.42497
  Tipo_FrasePM                 0.5976     0.4754   1.257  0.20877
  Tipo_FrasePL                 0.7260     0.4784   1.518  0.12912
  AutomaticoTO:Tipo_FraseFL   -0.9586     0.6706  -1.429  0.15288
  AutomaticoTO:Tipo_FrasePC   -0.6979     0.6751  -1.034  0.30126
  AutomaticoTO:Tipo_FrasePM   -0.8681     0.6698  -1.296  0.19493
  AutomaticoTO:Tipo_FrasePL   -1.6302     0.6726  -2.424  0.01536 *
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

car::Anova(m0, type=3)

  Analysis of Deviance Table (Type III Wald chisquare tests)

  Response: Aceptación
                          Chisq Df Pr(>Chisq)
  (Intercept)           0.2166  1   0.641644
  Automatico            8.7632  1   0.003074 **
  Tipo_Frase            2.6876  4   0.611394
  Automatico:Tipo_Frase 6.0332  4   0.196684
```

Ilustración 41. Modelo para el análisis estadístico

Posteriormente, se realizan las estimaciones basadas en el modelo (Ilustración 42, página siguiente) para calcular la probabilidad de aceptación según el tipo de tuit y su longitud. En la columna *prob* ya se puede apreciar, por un lado, que a nivel global los tuits originales tienen una mayor probabilidad de aceptación y que la única categoría según

longitud que supera en probabilidades a los tuits originales son los tuits formados por un párrafo de frases largas.

```
library(emmeans)
emmeans(m0, ~ Automatico | Tipo_Frase, type="response")

Tipo_Frase = FC:
 Automatico  prob     SE  df asymp.LCL asymp.UCL
 TA         0.461 0.0832 Inf     0.307     0.623
 TO         0.777 0.0589 Inf     0.642     0.872

Tipo_Frase = FL:
 Automatico  prob     SE  df asymp.LCL asymp.UCL
 TA         0.580 0.0830 Inf     0.415     0.730
 TO         0.684 0.0718 Inf     0.530     0.806

Tipo_Frase = PC:
 Automatico  prob     SE  df asymp.LCL asymp.UCL
 TA         0.557 0.0858 Inf     0.388     0.713
 TO         0.718 0.0673 Inf     0.570     0.830

Tipo_Frase = PM:
 Automatico  prob     SE  df asymp.LCL asymp.UCL
 TA         0.609 0.0812 Inf     0.444     0.752
 TO         0.727 0.0653 Inf     0.583     0.835

Tipo_Frase = PL:
 Automatico  prob     SE  df asymp.LCL asymp.UCL
 TA         0.639 0.0795 Inf     0.474     0.777
 TO         0.586 0.0799 Inf     0.426     0.729

Confidence level used: 0.95
Intervals are back-transformed from the logit scale
```

Ilustración 42. Estimaciones basadas en el modelo

Finalmente, se realizan las comparaciones basadas en el modelo para obtener el *Odds Ratio* que se encuentra en la primera columna de la Ilustración 43.

```
p0 <- pairs( emmeans(m0, ~ Automatico | Tipo_Frase, type="response") )
p0

 Tipo_Frase                          =                          FC:
  contrast   odds.ratio           SE     df   null   z.ratio   p.value
  TA  /  TO          0.245  0.116  Inf        1    -2.960    0.0031

 Tipo_Frase                          =                          FL:
  contrast   odds.ratio           SE     df   null   z.ratio   p.value
  TA  /  TO          0.639  0.303  Inf        1    -0.945    0.3446

 Tipo_Frase                          =                          PC:
  contrast   odds.ratio           SE     df   null   z.ratio   p.value
  TA  /  TO          0.493  0.236  Inf        1    -1.477    0.1396

 Tipo_Frase                          =                          PM:
  contrast   odds.ratio           SE     df   null   z.ratio   p.value
  TA  /  TO          0.584  0.275  Inf        1    -1.140    0.2542

 Tipo_Frase                          =                          PL:
  contrast   odds.ratio           SE     df   null   z.ratio   p.value
  TA  /  TO          1.251  0.594  Inf        1     0.472    0.6369

 Tests are performed on the log odds ratio scale
```

Ilustración 43. Comparaciones basadas en el modelo

Los valores de la columna *Odds Ratio* confirman lo observado en el punto anterior. Los tuits de frases cortas o párrafos formados por frases cortas tienen menos probabilidades de ser aceptados (0,245 para FC y 0,493 para PC) que los formados por frases largas o párrafos mixtos (0,639 para FL y 0,584 para PM), aunque en estos cuatro casos la TAN no superaría a los tuits originales. La tendencia sí se invierte en los párrafos formados por frases largas (valor de 1,251). En esta variable, la TAN tiene más posibilidad de ser aceptada que los textos originales. Podemos observar este cambio de dirección en la siguiente ilustración:

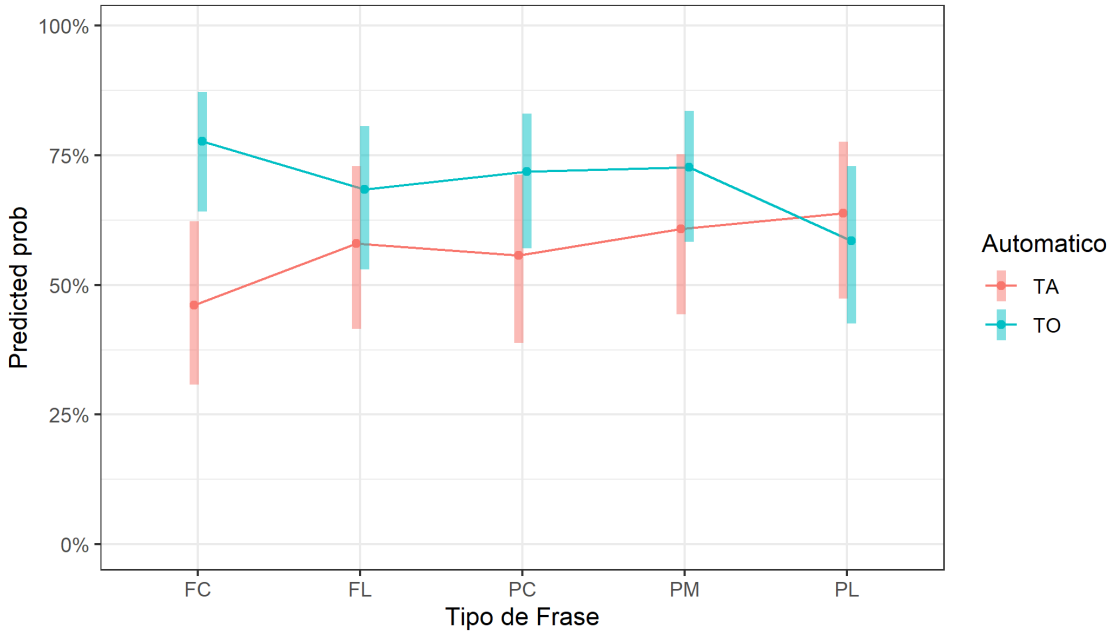


Ilustración 44. Comparaciones de probabilidades según la aceptación del tuit traducido y tuit original

Como podemos ver, los tuits originales siempre se sitúan por encima de los tuits traducidos, salvo en los párrafos largos. Aquí esta tendencia se invierta. Asimismo, también se puede apreciar cómo hay una mayor distancia entre los dos pares de los tuits formados por frases cortas y por párrafos de frases cortas.

4.4. Análisis de no inferioridad

Para llevar a cabo el análisis de no inferioridad, se modeliza la relación en términos de la aceptación de los tuits y las características de estos a través de un modelo lineal generalizado mixto con función de enlace *identity*. A continuación, podemos ver el código empleado.

```
m1 <- glmmPQL( Aceptación  ~ Automatico + Tipo_Frase +  Automatico : Tipo_Fra
se, random=list(ID=~1, Item=~1), family=binomial(link='identity'), data=bd3)
print(summary(m1),  correlation = FALSE)
```

```
  Linear mixed-effects model fit by maximum likelihood
    Data: bd3
    AIC BIC logLik
     NA  NA     NA

  Random effects:
   Formula: ~1 | ID
          (Intercept)
  StdDev:   0.1125966

   Formula: ~1 | Item %in% ID
           (Intercept)  Residual
  StdDev: 3.807198e-06 0.9835712

  Variance function:
   Structure: fixed weights
   Formula: ~invwt
  Fixed effects:  Aceptación ~ Automatico + Tipo_Frase + Automatico:Tipo_Fras
e
                                Value  Std.Error   DF    t-value p-value
  (Intercept)                0.3845701 0.02198307 4950  17.493924  0.0000
  AutomaticoTO               0.3335720 0.02766064 4950  12.059449  0.0000
  Tipo_FraseFL               0.1252825 0.03153297 4950   3.973063  0.0001
  Tipo_FrasePC               0.0881506 0.03388155 4950   2.601727  0.0093
  Tipo_FrasePM               0.1844331 0.03380470 4950   5.455842  0.0000
  Tipo_FrasePL               0.2957608 0.03029883 4950   9.761458  0.0000
  AutomaticoTO:Tipo_FraseFL -0.1368892 0.03998844 4950  -3.423220  0.0006
  AutomaticoTO:Tipo_FrasePC -0.0729627 0.04121386 4950  -1.770345  0.0767
  AutomaticoTO:Tipo_FrasePM -0.1441153 0.04057189 4950  -3.552098  0.0004
  AutomaticoTO:Tipo_FrasePL -0.4256167 0.04082212 4950 -10.426130  0.0000
   Correlation:
                             (Intr) AtmtTO Tp_FFL Tp_FPC Tp_FPM Tp_FPL ATO:T_F
F
  AutomaticoTO               -0.720
  Tipo_FraseFL               -0.628  0.502
  Tipo_FrasePC               -0.587  0.472  0.409
  Tipo_FrasePM               -0.595  0.486  0.412  0.397
  Tipo_FrasePL               -0.659  0.536  0.459  0.435  0.454
  AutomaticoTO:Tipo_FraseFL  0.500 -0.693 -0.791 -0.329 -0.339 -0.373
  AutomaticoTO:Tipo_FrasePC  0.485 -0.670 -0.336 -0.827 -0.339 -0.365  0.466
  AutomaticoTO:Tipo_FrasePM  0.500 -0.688 -0.344 -0.336 -0.846 -0.385  0.484
  AutomaticoTO:Tipo_FrasePL  0.496 -0.693 -0.346 -0.331 -0.351 -0.755  0.483
                             ATO:T_FPC ATO:T_FPM
  AutomaticoTO
  Tipo_FraseFL
  Tipo_FrasePC
```

Ilustración 45. Código empleado para el cálculo de la no inferioridad

A partir de este modelo, se obtienen las estimaciones en bruto que vemos en la Ilustración 46 de la página 127. Sobre estas estimaciones se aplicará, posteriormente, el margen de

no inferioridad, más concretamente, sobre el límite superior del intervalo de confianza (columna upper.CL).

```
emmeans(m1, ~ Automatico | Tipo_Frase, type="response")

 Tipo_Frase = FC:
  Automatico emmean     SE  df lower.CL upper.CL
  TA          0.385 0.0220 260    0.341    0.428
  TO          0.718 0.0193 260    0.680    0.756

 Tipo_Frase = FL:
  Automatico emmean     SE  df lower.CL upper.CL
  TA          0.510 0.0246 260    0.461    0.558
  TO          0.707 0.0180 260    0.671    0.742

 Tipo_Frase = PC:
  Automatico emmean     SE  df lower.CL upper.CL
  TA          0.473 0.0275 260    0.419    0.527
  TO          0.733 0.0165 260    0.701    0.766

 Tipo_Frase = PM:
  Automatico emmean     SE  df lower.CL upper.CL
  TA          0.569 0.0272 260    0.515    0.623
  TO          0.758 0.0145 260    0.730    0.787

 Tipo_Frase = PL:
  Automatico emmean     SE  df lower.CL upper.CL
  TA          0.680 0.0229 260    0.635    0.725
  TO          0.588 0.0209 260    0.547    0.629

 Degrees-of-freedom method: containment
 Confidence level used: 0.95
```

Ilustración 46. Estimaciones basadas en el modelo de no inferioridad

Decidimos establecer el margen de no inferioridad en el 95 % unilateral fijado en el 20 %, es decir, establecer que la tasa de aceptación de la traducción automática es superior al valor de la tasa de aceptación de los tuits originales menos 20 puntos porcentuales. La elección de este margen es crítica y representa uno de los principales desafíos en los ensayos de no inferioridad, dado que no existe un consenso universal. Las agencias reguladoras proporcionan algunas recomendaciones sobre cómo elegir el margen, pero no fijan ninguno. Molina (2020) declara que en la mayoría de los casos se escoge un margen del 30 %, mientras que autores como Althunian *et al*. (2017) hablan de un 5 % o Senn (2000) de un 10 %.

De forma práctica esto equivale a que el límite superior del intervalo de confianza al 90 % para la diferencia entre tasas (TA - TO) esté por encima del valor -0.2 (valor *upper.CL* en la Ilustración 47).

```
Tipo_Frase = FC:
 contrast estimate     SE   df lower.CL upper.CL
 TA - TO    -0.334 0.0277 4950  -0.3791   -0.288

Tipo_Frase = FL:
 contrast estimate     SE   df lower.CL upper.CL
 TA - TO    -0.197 0.0288 4950  -0.2441   -0.149

Tipo_Frase = PC:
 contrast estimate     SE   df lower.CL upper.CL
 TA - TO    -0.261 0.0306 4950  -0.3109   -0.210

Tipo_Frase = PM:
 contrast estimate     SE   df lower.CL upper.CL
 TA - TO    -0.189 0.0294 4950  -0.2379   -0.141

Tipo_Frase = PL:
 contrast estimate     SE   df lower.CL upper.CL
 TA - TO     0.092 0.0294 4950   0.0436    0.140
```

Ilustración 47. Cálculo de la no inferioridad a partir de las estimaciones del modelo

Por tanto, tal y como vemos en la columna de resultados *upper.CL*, solo podemos declarar la no inferioridad en los tuits formados por un párrafo de frases largas. Si bien, cabe destacar que tanto los párrafos mixtos como las frases largas se sitúan cerca del umbral fijado en -0.2. Por el contrario, y tal y como veníamos viendo en las exploraciones previas, las frases cortas tanto solas como en párrafos son las más alejadas del umbral. Esto indica un camino claro en el que trabajar para mejorar el motor.

5. Validación del método de evaluación

Para poder saber si nuestro motor cumple con los objetivos que nos habíamos propuesto, debemos también evaluar el instrumento de análisis en sí mismo. Por tanto, durante esta prueba piloto, se extrae información para confirmar si el instrumento es capaz de recopilar todos los datos que necesitamos para evaluar el motor y si todas nuestras preguntas se

contestan de la manera que pretendíamos. Por esta razón, también realizamos un análisis descriptivo del propio instrumento.

En primer lugar, miramos en profundidad el tipo de respuestas para detectar cualquier anomalía. Para ello, analizamos la tasa de aceptación global de todas las respuestas y los 5 mejores y peores en la tasa de acierto.

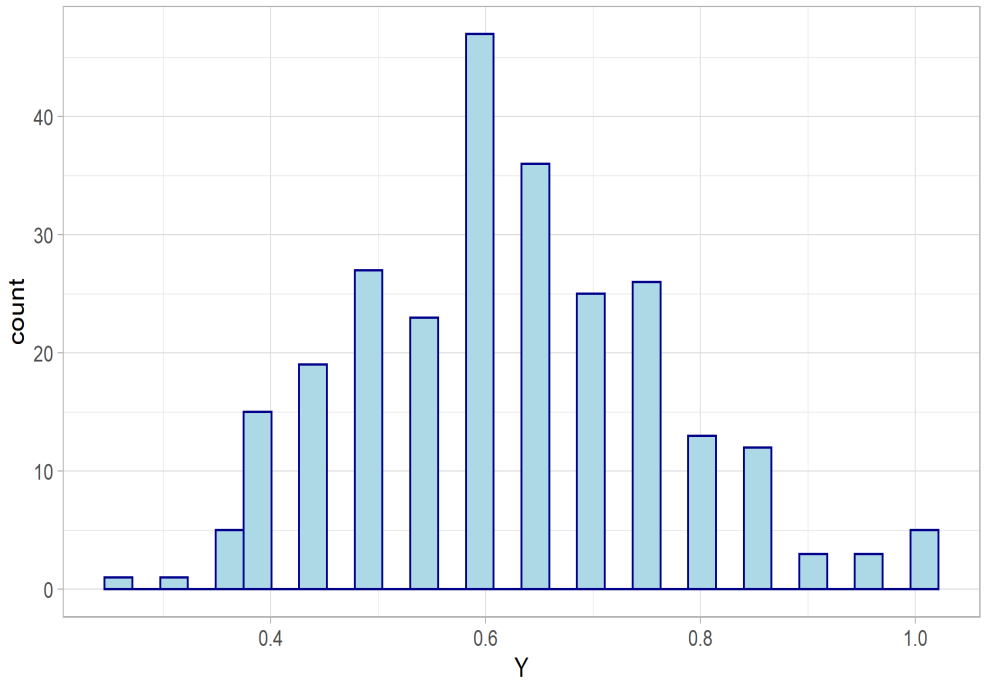


Ilustración 48. Histograma de respuestas según aceptación global

En la Ilustración 48, el vector Y va de 0, que representa la no aceptación de ningún tuit, a 1, que representa la aceptación de todos los tuits. Tal y como se puede apreciar, la mayoría de las respuestas se agrupan entre el 0,4 y el 0,8, lo que demuestra cierta heterogeneidad en las respuestas y confirma que la selección de los participantes fue válida.

No obstante, tal y como vemos en la
Tabla 6 de la página 130, hubo participantes que aceptaron todo como original y otros que rechazaron incluso tuits escritos originalmente en gallego.

ID	TA	TO	Acierto
60	25.0	37.5	31.25
70	25.0	37.5	31.25
63	16.7	50.0	33.35
146	16.7	50.0	33.35
16	27.3	44.4	35.85
259	100.0	71.4	85.70
157	83.3	92.9	88.10
192	83.3	92.9	88.10
90	85.7	92.3	89.00
169	100.0	85.7	92.85

Tabla 6. 5 mejores y 5 peores según la tasa de acierto

Cabe recordar que los cuestionarios se distribuyeron a través de Twitter dentro de comunidades especialmente activas en el uso y promoción del gallego, por lo que el perfil de los participantes se correspondería con el de un hablante que conscientemente elige utilizar como lengua vehicular el gallego en su día a día. En consecuencia, se podría argumentar que estos participantes son expertos en la lengua y que tienden a buscar la corrección y estilo lingüístico, algo que no siempre es posible debido a las restricciones de Twitter.

Por último, completamos nuestro análisis con la evaluación del perfil de las respuestas según las características del tuit, es decir, analizamos la aceptación de los participantes en función de si era un tuit original o traducido y de la longitud de estos (véase Tabla 7 de la página 131).

	Non	Si	p-value	N
	N=1965	*N=3255*		
Origen:			<0.001	5220
TA	958 (47.7%)	1049 (52.3%)		
TO	1007 (31.3%)	2206 (68.7%)		
Tipo_Frase:			<0.001	5220
FC	468 (45.4%)	563 (54.6%)		
FL	419 (38.4%)	671 (61.6%)		
PC	371 (35.9%)	661 (64.1%)		
PM	329 (30.8%)	740 (69.2%)		
PL	378 (37.9%)	620 (62.1%)		
Tipo_Automatico:			<0.001	5220
FC-TA	315 (61.9%)	194 (38.1%)		
FC-TO	153 (29.3%)	369 (70.7%)		
FL-TA	207 (49.2%)	214 (50.8%)		
FL-TO	212 (31.7%)	457 (68.3%)		
PC-TA	172 (51.8%)	160 (48.2%)		
PC-TO	199 (28.4%)	501 (71.6%)		
PL-TA	125 (31.0%)	278 (69.0%)		
PL-TO	253 (42.5%)	342 (57.5%)		
PM-TA	139 (40.6%)	203 (59.4%)		
PM-TO	190 (26.1%)	537 (73.9%)		

Tabla 7. Resumen descriptivo de las variables

Tal y como se aprecia en la Tabla 7, los *p*-valores son significativos por lo que prueban que el instrumento es correcto. También constatamos diferencias entre los participantes, por ejemplo, el porcentaje de aceptación de la variable «FC-TA» de la tabla es de un 38,1 %, muy alejado del resto de porcentajes de las otras categorías. Esta gran variación

se podría haber inferido si en la recolección de los datos se hubiesen incluido más preguntas demoscópicas para así conocer mejor al participante y evaluar las tasas intrapersonales o cualquier otro mecanismo que fijase objetivamente las características discursivas evaluadas. Es llamativo el porcentaje de tuits originales que no se han percibido correctamente como tales, por ejemplo, en la variable «PL-TO» de la tabla con un 57,5 %, el valor más bajo entre los tuits originalmente escritos en gallego. Este fenómeno puede estar provocado por el cuestionario, es decir, que la propia herramienta haya llevado al participante a identificar estilos correctos pero inusuales como TA, cuando de hecho, estaban leyendo textos originales.

6. Consideraciones finales

En este apartado, hemos descrito el proceso de elaboración y análisis de la prueba piloto con el doble objetivo de evaluar el motor entrenado y el propio instrumento de evaluación basado en el principio de no inferioridad.

Uno de los principales hallazgos es que los tuits escritos originalmente en gallego no siempre se perciben como tales. A partir de aquí, al calcular la no inferioridad con nuestro modelo estadístico, encontramos información clara con valor significativo de los puntos fuertes y débiles del motor. Las estimaciones basadas en el modelo nos indican dónde tenemos que mejorar el motor para alcanzar la no inferioridad en todos los tipos de tuits: en las frases cortas, tanto solas como en párrafos. Por otro lado, el sistema usado para entrenar la primera versión del motor no permitía ajustar la configuración de la longitud de los segmentos en el corpus, lo que podría haber penalizado también las frases cortas. Por ello, el motor se debe entrenar de nuevo con una mayor muestra de segmentos cortos para mejorar su rendimiento y con otra tecnología y, posteriormente, evaluarlo siguiendo métricas automáticas, humanas y de no inferioridad.

Por otro lado, cabe señalar que en las oraciones largas la primera versión de nuestro motor no es inferior a los tuits redactados originalmente en gallego. Este dato resulta extremadamente interesante ya que prueba que nuestro corpus es lo suficientemente auténtico y natural para los participantes en este tipo específico de frases. Más aún, si tenemos en cuenta que el participante estaba evaluando el resultado directo del motor sin poseditar.

En cuanto al método en sí mismo, demostramos que, aunque las evaluaciones de no inferioridad son más propias de los estudios médicos, se pueden aplicar también en los estudios de traducción automática para extraer las percepciones de los usuarios finales. Con este método, nos centramos en identificar las diferencias en la percepción cuando se comparan tuits traducidos automáticamente con tuits escritos originalmente en gallego. Sin embargo, pudimos detectar ciertos problemas que tendremos que subsanar en los cuestionarios finales cuando el motor se reentrene. Por un lado, debemos introducir algunas preguntas de control para detectar un nivel lingüístico básico de gallego y así excluir aquellos participantes que no lo cumplan. Por otro lado, debemos también introducir más contexto y cohesión en la comprensión de los tuits, por lo que decidimos incluir también hilos y no únicamente tuits. Finalmente, también debemos controlar que el número de respuestas en cada uno de los cuestionarios sea similar para evitar que una distribución desproporcionada influenciara en el análisis.

CAPÍTULO IX: Entrenamiento del motor: prueba final

En este capítulo, se aborda el procedimiento seguido para entrenar de nuevo el motor tras la baja aceptabilidad obtenida en las frases cortas durante la prueba piloto descrita en el capítulo anterior. En específico, en este capítulo se describen las modificaciones implementadas durante la segunda fase de entrenamiento del motor de TA. Dicha fase se caracteriza por la introducción de cambios significativos en dos aspectos fundamentales del diseño original del motor: la cantidad de datos del corpus específico empleado y la tecnología empleada.

Cabe recordar que uno de los objetivos principales de esta investigación es la creación de un motor de TAN apto para el gallego y las redes sociales. Nuestra hipótesis de partida es que se puede conseguir una calidad alta. En este capítulo se conseguirá, por tanto, alcanzar nuestro objetivo y tratar nuestra hipótesis.

Se puede consultar el modelo del motor final entrenado en la plataforma Github[20]. Para comodidad del usuario final, en la plataforma MutNMT se puede traducir directamente con la versión JoeyNMT (GalNMT2) del motor reentrenado[21].

1. Ampliación del corpus específico

Tras el análisis de los resultados obtenidos en la prueba piloto descrita en el capítulo anterior, pudimos concluir que nuestro corpus específico no permitía realizar un buen entrenamiento en frases cortas. Por ello, para intentar usar en el entrenamiento un corpus con una mayor representación de frases cortas, se decide mantener el corpus *Paracrawl* como base genérica, pero ampliar el corpus proveniente de Twitter. Inicialmente, nuestro corpus de Twitter comprendía apenas alrededor de 70,000 oraciones únicas. Para la construcción de un corpus de Twitter de mayor envergadura, se seleccionan cuentas adicionales de instituciones gallegas (ver Tabla 8), tales como aquellas asociadas a la

[20] Disponible en: https://github.com/mariadocampo/GalNMT [Fecha de consulta: 14 de septiembre de 2023]
[21] Disponible en: https://ntradumatica.uab.cat/translate/ [Fecha de consulta: 14 de septiembre de 2023]

televisión y radio oficiales de Galicia, cuentas destinadas a la promoción del gallego, cuentas de revistas de divulgación y cuentas de pódcast en gallego. Estas 18 cuentas se eligen específicamente siguiendo los mismos criterios que en el estudio previo, es decir, cuentas con un uso adecuado de la gramática y la ortografía, así como expresiones comunes y naturales.

Cuenta de Twitter	**Número de tuits**
@uvigo	11507
@UDC_gal	10258
@UniversidadeUSC	7875
@AcademiaGalega	6362
@PortalPalabras	5543
@SXPL	4165
@Falaredes	8887
@culturagaleg	28252
@consellocultura	5486
@biosbardia	4191
@EdGalaxia	13518
@podgalego	5644
@comochodigo	362
@ctnl	9994
@NeoFalantes	362
@GalegoTwitch	13223
@diariocultural_	13545
@RadioGalega	81604
@TVGalicia	144741
@DigochoEuTVG	738
@IGE_Estatistica	27131
@Valedordopobo	3385
@Fegamp	3743
@Par_Gal	17258

Tabla 8. Cuentas y número de tuits utilizados para ampliar el corpus específico

Seguimos los mismos pasos y *scripts* descritos en el apartado 1.2 Creación del corpus específico del capítulo IV de la página 98 para extraer, limpiar y procesar el nuevo corpus. El corpus bilingüe depurado contiene 262 785 oraciones únicas y 5 448 375 palabras. Entrenamos nuestro motor con el corpus *Paracrawl* genérico español - gallego (1 879 651 oraciones y 44 626 394 palabras) y este corpus específico.

	Motor prueba piloto	**Motor evaluación final**
Corpus Twitter	69 713	262 785
Corpus *Paracrawl*	1 879 651	1 879 651

Tabla 9. Comparativa de datos usados para el entrenamiento en el motor de la prueba piloto y el de la evaluación final

2. Elección de la tecnología TAN

Tras el entrenamiento del motor para la prueba piloto descrito en el capítulo V, se realiza un segundo entrenamiento con otra tecnología de software libre que permite controlar un poco más los parámetros de entrenamiento. La tecnología en este caso es Marian[22], versión 1.12.0 (Junczys-Dowmunt *et al.*, 2018). Marian es un marco de traducción automática neural escrito en C++ con dependencias mínimas. El equipo de Microsoft Translator es el principal desarrollador, aunque muchos académicos (principalmente de la universidad de Edimburgo y, en el pasado, de la universidad de Poznań) y otros contribuidores comerciales ayudan en su desarrollo. Actualmente, es la tecnología que está detrás del traductor neuronal Microsoft Translator y se emplea en numerosas empresas, organizaciones y proyectos de investigación.

Estas son algunas de sus funcionalidades:

- Implementación pura y eficiente en C++
- Entrenamiento rápido multi-GPU y traducción GPU/CPU
- Arquitecturas NMT de última generación: *deep RNN* y *transformer*
- Licencia de libre acceso permisiva (MIT)

[22] Disponible en: https://github.com/marian-nmt/marian/releases/tag/1.12.0. [Fecha de consulta: 13 de septiembre de 2023]

Como vemos Marian ofrece más robustez a la hora de entrenar y permite un mejor control de los parámetros de entrenamiento y, por eso, la elegimos para nuestro segundo entrenamiento del motor, tal y como veremos en las conclusiones de la prueba piloto.

3. Proceso de entrenamiento

Para el entrenamiento con Marian, primero teníamos que disponer de una GPU con la capacidad suficiente para ejecutar entrenamientos neuronales, con un sistema operativo Ubuntu y con Marian compilado y listo para usar. La empresa de traducción CPSL facilitó la GPU en la que poder realizar el entrenamiento con Marian, ya que disponían de esta máquina previamente configurada para ello.

Otra de las decisiones que tomamos es el tipo de configuración. Marian proporciona la opción *--task*, que ofrece una manera rápida de configurar un entrenamiento de un tipo determinado. La lista de configuraciones predefinidas incluye:

- *best-deep*: la arquitectura RNN BiDeep propuesta por Miceli Barone *et al*. (2017)
- *transformer-base* y *transformer-big*: arquitecturas y configuraciones de entrenamiento para los modelos *transformer base* y *transformer big*, respectivamente, presentados en Vaswani *et al*. (2019)
- *transformer-base-prenorm* y *transformer-big-prenorm*: variantes de dos modelos *transformer* con la normalización de la capa se ejecuta como el primer paso de preprocesamiento por bloques.

En nuestro caso, el modelo que mejor se ajusta a nuestras necesidades y que mejores referencias había obtenido en otras investigaciones similares a la nuestra (Guerberof *et al*., 2022) era el *transformer-big*.

Otro de los elementos que queríamos controlar son las métricas de validación y en qué momento se paraba el entrenamiento. Normalmente, se detiene un entrenamiento cuando no se consiguen mejoras en los resultados de la métrica o métricas de validación durante un determinado número de validaciones (lo que se conoce como *early stopping*). Marian ofrece varias métricas de validación: *cross-entropy* (métrica que analiza a nivel de segmento cómo de bueno es el rendimiento de un modelo con un número del 1 al 0, siendo este último un modelo sin errores), *ce-mean-words* (lo mismo que el anterior, pero a nivel de palabra), *valid-script* (ejecuta un *script* propio que debe devolver un número),

Translation (ejecuta también un *script* propio que debe devolver un número), *bleu* (calcula la métrica BLEU en el set de validación en bruto) y *bleu-detok* (calcula la métrica BLEU en el set de validación posprocesado). En nuestro caso, empleamos *cross-entropy* y *bleu-detok*. Por su parte, el *early stopping* es la técnica habitual para decidir cuándo detener un entrenamiento basada en una heurística que utiliza el conjunto de validación. Por defecto se usa una paciencia de 10. Esto quiere decir que el entrenamiento se parará cuando pasen 10 procesos de validación iterativos en los que no se consigan mejoras. Si se utiliza más de una métrica en la validación por defecto se detiene el entrenamiento cuando cualquiera de estas métricas llega al valor de *early-stopping*. No obstante, en nuestro caso, indicamos que se detuviera cuando todas las métricas llegasen al valor de *early-stopping* poniendo el valor del parámetro *--early-stopping-on* en *all*. Además, también controlamos dos parámetros adicionales relacionados con *early-stopping*:

- *overwrite*: por defecto Marian guarda todos los modelos para cada paso de validación. Si ponemos este parámetro en *true* borrará los modelos anteriores y mantendrá únicamente el último (o el mejor, dependiendo de la siguiente opción).
- *keep-best*: se emplea junto a la opción anterior. Si está en true guarda el mejor modelo.

A continuación, podemos ver un ejemplo de nuestro archivo de configuración:

```
authors: false
cite: false
task: transformer-big
train-sets:
  - train.sp.es
  - train.sp.gl
valid-sets:
  - val.sp.es
  - val.sp.gl
guided-alignment: train.sp.es.gl.align
vocabs:
  - vocab-es.yml
  - vocab-gl.yml
early-stopping: 10
keep-best: true
overwrite: true
dim-vocabs:
  - 64000
                                -                                64000
workspace: 10000
valid-metrics:
    - cross-entropy
    - bleu-detok
```

Ilustración 49. Parámetros de configuración del entrenamiento en Marian

Los parámetros como los mostrados en la Ilustración 49 se guardan en un archivo de configuración, que debe estar en la misma ruta en la que están todos los corpus preprocesados previamente y a continuación se ejecuta el siguiente comando:

```
./marian -c archivodeconfiguración.yml
```

Ilustración 50. Comando para iniciar el entrenamiento en Marian

La duración del entrenamiento dependerá de la capacidad de la GPU, del tamaño del corpus y de las métricas empleadas. En nuestro caso, el motor tardó dos días en entrenarse en la GPU con tarjeta gráfica NVIDIA, 32 gigas de memoria RAM y procesador Intel Core I7-9700G9 @3.0GHz. Una vez entrenado, los módulos MTUOC permiten montar un servidor Marian con el que poder comunicarse a través de la herramienta de traducción asistida SDL Trados Studio y traducir directamente desde allí.

Una vez entrenado el motor de nuevo, es necesario evaluarlo de nuevo. Para ofrecer una visión completa de la evaluación, hemos unido en el capítulo siguiente la evaluación automática, la evaluación humana siguiendo el marco MQM-DQF y la evaluación de la percepción de los usuarios siguiente el principio de no inferioridad.

CAPÍTULO X: Evaluación del motor de la prueba final

La información obtenida a partir de la prueba piloto nos permite perfeccionar el procedimiento seguido para evaluar finalmente el motor reentrenado gracias a la información extraída en la prueba piloto. En esta segunda evaluación, combinamos el enfoque tradicional de evaluación automática y humana de la TAN con la evaluación de no inferioridad. El objetivo principal de este capítulo, por tanto, es presentar y discutir los resultados obtenidos en las tres evaluaciones y correlacionar los datos para demostrar la calidad y utilidad del motor, además de consolidar el sistema de evaluación sin *gold standard* a través del método de no inferioridad.

En este proceso, primero se lleva a cabo una evaluación automática con la métrica BLEU en el apartado 1. Posteriormente en el apartado 2, se encarga a tres traductores gallegos con más de 10 años de experiencia la evaluación humana de los tuits traducidos automáticamente siguiendo el marco MQM-DQF de clasificación de errores. Finalmente en el apartado 3, se aplica la estrategia de evaluación de la no inferioridad con el objetivo de relacionar una mala percepción del usuario con errores graves del motor detectados en la evaluación humana. Se introducen en el instrumento de recogida de datos los cambios necesarios constatados en la prueba piloto. Para ello, primero se crea una nueva muestra de tuits escritos originalmente en gallego y tuits escritos en castellano que se tradujeron utilizando el motor de traducción automática desarrollado. Como se explicará más en detalle a continuación, los tuits se vuelven a clasificar según su longitud y esta vez se incluyen hilos para intentar corregir la falta de contexto detectada en la prueba piloto. Posteriormente, se diseñan distintos cuestionarios y se distribuyen entre una muestra diferente de usuarios de Twitter gallegos para extraer su percepción. Los datos se analizan de nuevo con la ayuda del servicio de estadística de la UAB.

Los resultados obtenidos durante esta triple evaluación proporcionan una visión detallada sobre la calidad de las traducciones generadas por el motor final y la solidez del método de evaluación.

1. Evaluación automática

Tal y como se explicó en el apartado 1.2.4 referente al proceso de preprocesamiento del corpus, este se divide en *train*, *val* y *eval*. La partición *eval* está compuesta exclusivamente de segmentos procedentes del corpus específico de Twitter. El motor reentrenado se evalúa automáticamente contra esta partición para evaluación y obtiene un porcentaje BLEU del 85 %. Recordemos que cuanto más se parezca el resultado directo de la traducción automática a la traducción humana de referencia, mayor será el porcentaje BLEU. Si comparamos el BLEU obtenido con motores con lenguas de características similares, por ejemplo, el motor ES/EN-GL de Ortega *et al.* (2022: 93) con un porcentaje en torno al 50 %; el motor *Nós* de Gamallo *et al*. 2023 con un porcentaje del 79,6 %; el motor EN-GA de Lankford (2022: 9) con un porcentaje del 60,5 %; o el motor ES-CA de Rapp (2021: 295) con un porcentaje en torno al 79 %, constatamos que nuestro motor obtiene un BLEU por encima de la bibliografía consulta.

A su vez, si se compara este BLEU con el obtenido durante el primer entrenamiento (véase Tabla 11), se puede apreciar que el porcentaje ha mejorado un 15 %, lo que supone un incremento sustancial.

Tipo de motor	BLEU
JoeyNMT	70,63 %
Marian	85 %

Tabla 10. Comparativa de la evaluación automática de los dos motores entrenados

Por tanto, desde la perspectiva de la evaluación automática, el primer objetivo de investigación, esto es, crear un motor en una combinación de idiomas con pocos recursos, se ha conseguido.

A pesar del buen resultado obtenido, cabe recordar que las métricas automáticas, en especial, BLEU no siempre se correlacionan con las evaluaciones humanas, ya que son incapaces de reconocer y de capturar la similitud semántica más allá del nivel léxico (Rei *et al.*, 2020: 2692). La evaluación humana, por su lado, captura mucho más que la similitud semántica, por eso decidimos realizar una clasificación de errores siguiendo el marco MQM-DQF (Lommel *et al.*, 2018) y completarla con la evaluación de no inferioridad.

2. Evaluación humana (MQM-DQF)

Para poder evaluar el motor y garantizar la información de los cuestionarios descrita en el apartado 3, realizamos una evaluación humana de categorización de errores basada en el marco MQM-DQF (Lommel *et al.*, 2018; Görög, 2014; Popovi´c, 2018). Encargamos a tres traductores profesionales con más de diez años de experiencia que realizasen esta evaluación mediante la plataforma en línea ContentQuo[23]. Esta plataforma está especialmente pensada para la evaluación de la calidad de la traducción y proporciona distintos flujos de trabajo y plantillas predefinidas con estándares de calidad, como, por ejemplo, la empleada, DQF. El objetivo de realizar ahora esta evaluación es encontrar un vínculo entre los errores reportados y una actitud negativa hacia los tuits traducidos automáticamente, de manera similar a la metodología empleada por Guerberof *et al.*, 2022 y Bhardwaj *et al.*, 2020.

La selección de los traductores se realiza mediante el portal Proz.com filtrando por más de 10 de experiencia en la combinación castellano-gallego, experiencia en traducción automática y lengua materna gallego. Esta evaluación profesional es remunerada y los traductores dedicaron de una hora a dos horas para llevarla a cabo. Se les solicita a los revisores revisar la traducción automática sin poseditar de 30 tuits (890 palabras). Estos tuits traducidos con nuestro motor son los mismos tuits que se presentan en los cuestionarios de los usuarios finales. Antes de empezar la evaluación en ContentQuo se ofrece una pequeña formación en línea mediante el programa Teams para contextualizar la tarea dentro de esta tesis de investigación, explicar el encargo y enseñar cómo trabajar dentro de la herramienta. Se les da a los traductores la instrucción exacta de leer cada segmento traducido, detectar cualquier error de traducción y clasificarlo según el marco propuesto dentro de la herramienta. No se les solicita poseditar el texto, pero sí se les pide una valoración general sobre la calidad que percibían.

2.1. Trabajo en la plataforma

ContentQuo muestra al usuario el texto origen y la traducción en dos columnas, de manera similar a como se ve en una herramienta de traducción asistida (ver Ilustración 51 de la página 143).

[23] Disponible en: https://www.contentquo.com/. Fecha de consulta: 18 de agosto de 2023.

Evaluación humana (MQM-DQF)

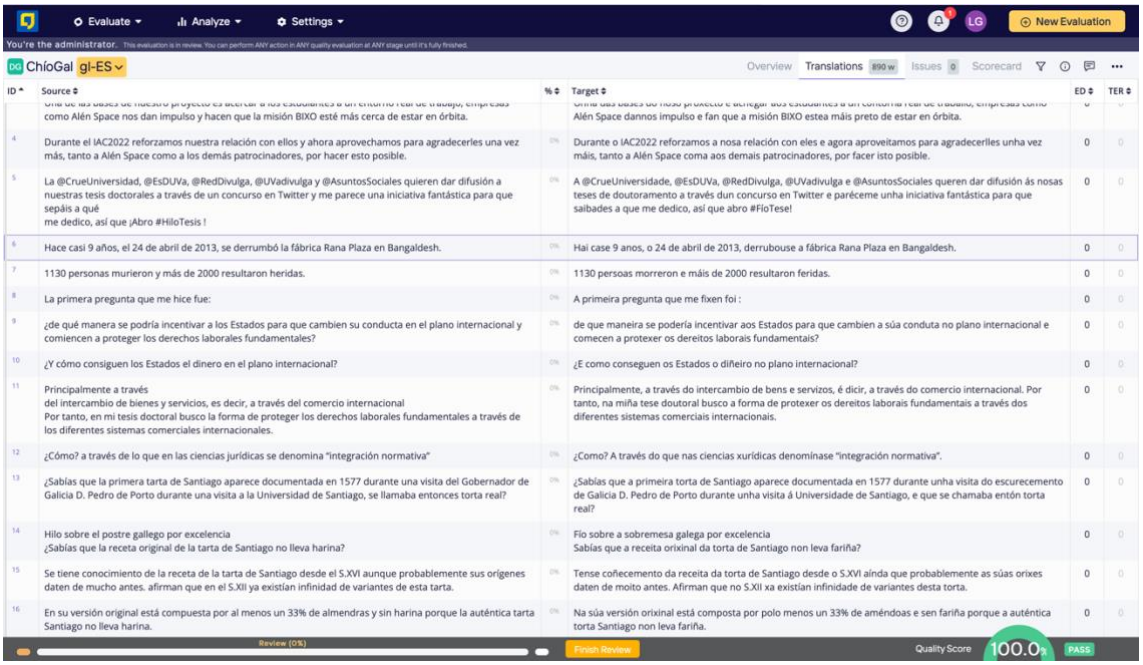


Ilustración 51. Pantalla principal de ContentQuo

Los traductores, por tanto, no ven el texto en formato tuit, sino que ven los tuits divididos en segmentos. Cuando el usuario detecta un error, debe seleccionar la palabra y al hacer clic sobre ella se mostrará una ventana emergente (véase Ilustración 52) en la que seleccionar en un menú desplegable los errores definidos en el marco DQF (véase Ilustración 53 de la página 144), los distintos niveles de gravedad y un campo de texto para describir el error si el usuario lo considera necesario.

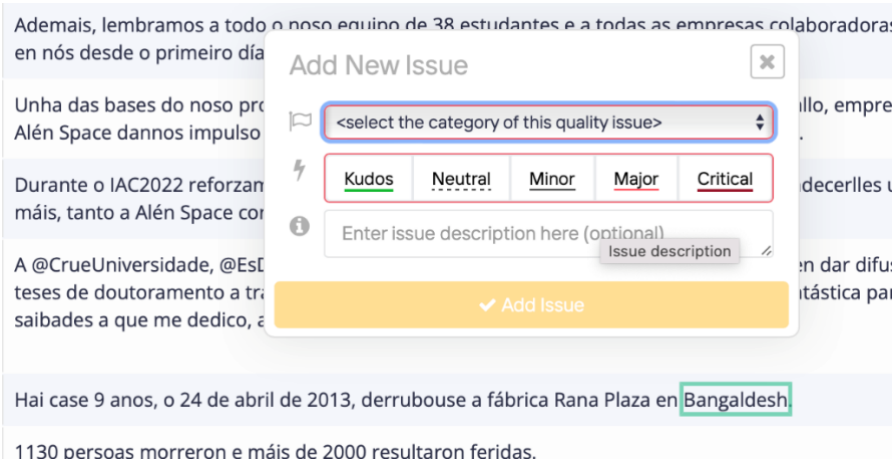


Ilustración 52. Ventana emergente para categorizar el error detectado en la evaluación

Como podemos apreciar en la siguiente imagen, la lista despegable para seleccionar el tipo de error incluye tanto las ocho clasificaciones principales como las subcategorías dentro de las mismas.

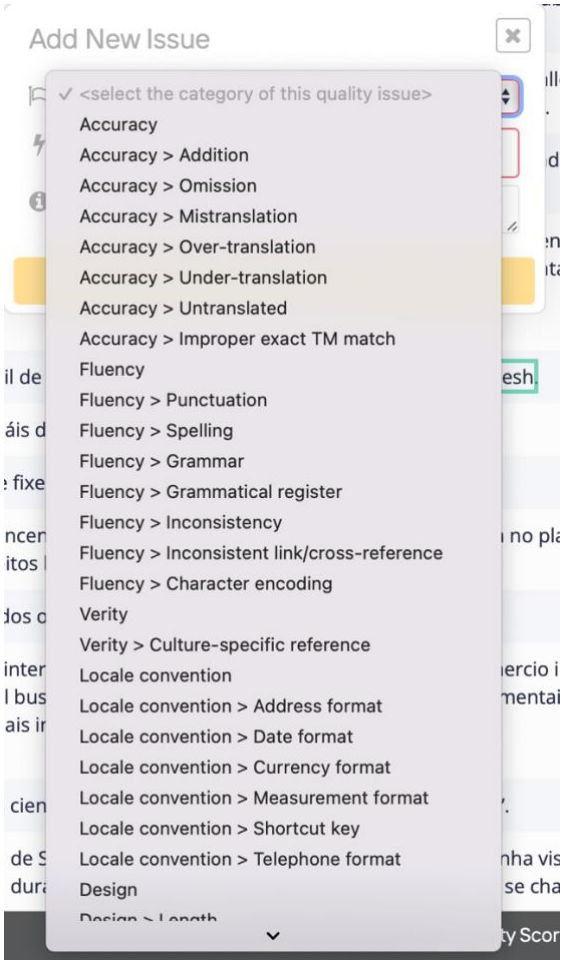


Ilustración 53. Lista de errores predefinidos en el marco MQM-DQF

El marco MQM-DQF establece los siguientes errores (Lommel *et al.*, 2018):

- Precisión (*accuracy* en inglés), cuando el texto de destino no refleja con exactitud el texto de partida. Dentro esta categoría, tenemos las siguientes 7 subcategorías:
  - Adición (*addition*), cuando hay elementos en el texto de destino que no están presentes en el texto de partida.
  - Omisión (*omission*), cuando elementos del texto de partida no están en el texto de destino.
  - Traducción incorrecta (*mistranslation*).
  - «Sobretraducción» (*over-translation*), cuando la información del texto de destino es más específica que el de partida.

  - «Subtraducción» (*under-translation*), cuando la información del texto de destino es menos específica que el de partida.
  - Contenido no traducido (*untranslated*).
  - Coincidencia exacta de la memoria de traducción incorrecta (*improper exact TM match*).
- Fluidez (*fluency* en inglés), cuando los errores están relacionados con la forma o el contenido del texto de destino y afectan negativamente a la comprensión del texto. En esta categoría hay siete subcategorías:
  - Puntuación (*punctuation*).
  - Ortografía (*spelling*) .
  - Gramática (*grammar*).
  - Registro lingüístico (*grammatical register*).
  - Inconsistencia (*inconsistency*).
  - Enlace/referencia cruzada (*link/cross-reference*).
  - Codificación de caracteres (*character encoding*).
- Terminología (*terminology* en inglés), cuando el uso de términos es distinto al esperado para el ámbito o se especifica de otro modo. Hay dos subcategorías:
  - Inconsistente con la base de datos terminológica (*inconsistent with termbase*), que en nuestro caso no era aplicable.
  - Uso incoherente de la terminología (inconsistent use of terminology).
- Estilo (*style* en inglés) de la redacción, como por ejemplo el empleo de un estilo serio en el texto de destino donde lo que requiere es un estilo ligero y humorístico. Hay cinco subcategorías:
  - Estilo extraño (*awkward*).
  - Estilo empresarial (*company style*), que en nuestro caso no aplicaba.
  - Estilo inconsistente (*inconsistent style*).
  - Estilo de terceros (*third-party style),* cuando el texto incumple con una guía de estilo diseñada por terceros. En nuestro caso no aplicaba.
  - Estilo no idiomático (*unidiomatic*).
- Diseño (*design* en inglés), cuando el texto está incorrectamente formateado o maquetado. Hay cinco subcategorías: longitud (*length*), formato local (*local formatting*), marcado (*markup*), texto perdido (*missing text*),

truncamiento/expansión de texto (*truncation/text expansion*). En nuestro caso, esta categoría de errores no aplicaba.

- Convención local (*locale convention* en inglés), cuando el texto de destino no se adapta a las tradiciones culturales específicas de la región y no cumple con los requisitos para la localización de esta región. Esta categoría se divide en seis subcategorías:
  - Formato de dirección (*address format*).
  - Formato de fecha (*date format*).
  - Formato de moneda (*currency format*).
  - Formato de medida (*measurement format*).
  - Tecla de atajo (*shortcut key*).
  - Formato de teléfono (*telephone format*).
- Veracidad (*verity* en inglés), cuando en el texto de destino hay afirmaciones que contradicen el contenido del texto de partida. Para esta categoría solo hay una subcategoría; referencia específica de la cultura (*culture-specific reference*).
- Otro, en esta categoría se engloban errores que no se han incluido en las categorías anteriores.

Además de clasificar los errores, el marco MQM-DQF también define la severidad del error en tres categorías: crítico, grave o leve, que se describirán a continuación:

- Error crítico: errores que pueden provocar consecuencias sanitarias, de seguridad, legales o financieras, que incumplan las normas de uso geopolíticas, que perjudiquen la reputación de la organización o entidad, que provoquen la suspensión de la aplicación informática o que modifiquen de forma negativa la funcionalidad de un producto o servicio, incluso informaciones que puedan considerarse ofensivas.
- Error grave: errores que pueden confundir o engañar al usuario o dificultar el uso correcto del producto/servicio debido a que en el texto de destino hay un cambio significativo en el contenido, o errores que aparecen en una parte visible o importante del contenido.
- Error leve: errores que no provocan directamente la pérdida de significado y que no confundirían o engañarían al usuario, pero que sí son detectados por los lectores. La influencia más negativa que pueden tener sería la disminución

de la calidad estilística, la fluidez o la claridad, o harían el contenido del texto de partida menos atractivo que el texto de origen.

- Error neutro: cambio preferencial del revisor.

Aunque no está pensado en el marco *per se*, ContentQuo ofrece también la posibilidad de utilizar la categoría *Kudos* como preferencial para marca algo excepcionalmente bueno en la traducción.

Una vez están todos los errores categorizados, ContentQuo realiza la siguiente operación para calcular el porcentaje de calidad:

QUALITYSCORE = 100 - 100 * (ERRORPOINTS / WORDCOUNT)

Ilustración 54. Fórmula empleada por ContentQuo para calcular el porcentaje de calidad

La fórmula (véase Ilustración 54) en sí misma calcula la calidad en términos de porcentaje y se basa en la relación entre los errores (*errorpoints*) y la cantidad de palabras (*wordcount*). Primero, se calcula la proporción de errores respecto a la cantidad total de palabras. Luego, se multiplica esta proporción por 100 para convertirla en un porcentaje. Esto da como resultado el porcentaje de errores en el texto en términos de calidad. Finalmente, se resta este porcentaje del valor 100 para obtener la puntuación de calidad final. Esto se hace para expresar la calidad en términos de puntaje, donde un puntaje más alto indica una mejor calidad.

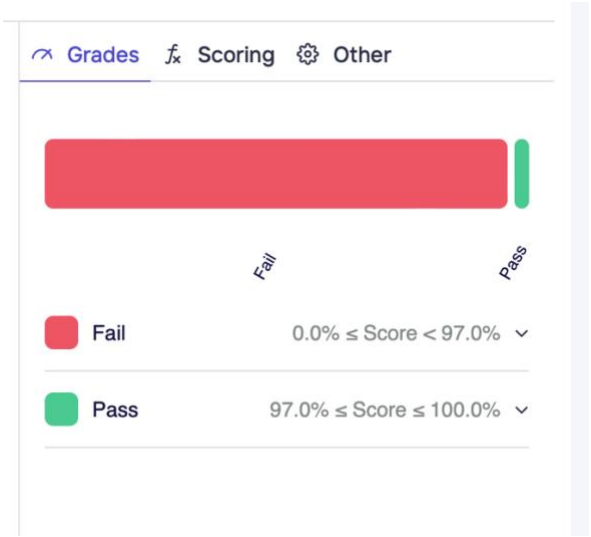


Ilustración 55. Umbral de éxito en el marco MQM-DQF

Cabe señalar que lo estándar en la industria es que el porcentaje de éxito se sitúe en el 97 % (véase Ilustración 55). Aunque en nuestra evaluación también obtendremos este porcentaje, es de mayor interés para el objetivo de esta investigación el tipo de errores detectados y su gravedad para poder correlacionarlos después.

2.2. Análisis de resultados

Para poder analizar los resultados de las distintas evaluaciones y presentar un análisis conjunto, se realiza una media del porcentaje final de las tres evaluaciones. No obstante, no hubo consenso en la clasificación de los errores, por lo que se contabilizan todos, excepto si se repite. Aunque como hemos explicado anteriormente la media global obtenida no es relevante para nuestro análisis, sí podemos destacar que se obtuvo un resultado de calidad del 94,55 %. Si bien estamos más interesadas en la lista de errores, consideramos que se trata de un buen resultado, muy cercano al aprobado de la industria, especialmente si tenemos en cuenta que la evaluación se ha realizado sobre el resultado directo del motor sin poseditar.

Categoría de error	**Neutral**	**Leve**	**Grave**	**Crítico**
Veracidad	0	0	0	0
Terminología	0	2	2	0
Estilo	1	2	0	0
Otro	3	3	0	0
Convención local	0	0	0	0
Fluidez	0	14	2	0
Diseño	0	0	0	0
Precisión	2	4	3	0

Tabla 11. Lista de tipos de errores según el marco MQM-DQF.

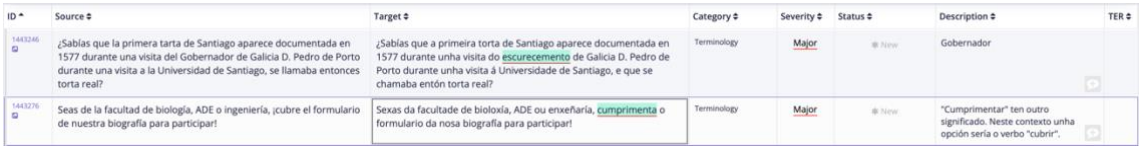
En cuanto a los tipos de errores, tal y como se puede apreciar en la Tabla 11, no se encuentran errores en las categorías veracidad, convención local y diseño. Por otro lado, se encuentran algunos errores en las categorías de estilo y otros. Tal y como era de esperar, la mayoría de los errores se concentran en las categorías de fluidez, precisión y

terminología. Dentro de la categoría de fluidez, se encuentran errores de gramática, puntuación y ortografía, mientras que en precisión la mayoría de los errores son traducciones erróneas y «sobretraducciones». En cuanto a su severidad, no se contabilizan errores críticos y sólo seis se clasifican como neutros. La mayoría de los errores, el 65 % del total de errores, están categorizados como leves. Por otra parte, se encuentran siete errores graves que describiremos a continuación.

2.2.1. Errores graves de terminología

A continuación, se presentan los errores de terminología detectados, así como el nivel de acuerdo entre revisores al marcar la severidad de estos. También se indica aquí a qué tipo de tuit corresponden. Por un lado, tenemos la traducción incorrecta de gobernador por *escurecemento*, que en castellano significa la acción y efecto de oscurecer. Este segmento pertenece a un hilo y una de las posibles interpretaciones de esta mala traducción puede ser que el término gobernador no es muy común en el corpus de entrenamiento. Todos los revisores lo han marcado como grave.

Por otro lado, se ha traducido el verbo cubrir por *cumprimentar*, que en castellano significa agasajar. Este error se ha encontrado en un tuit de un párrafo formado por frases cortas y se trata de un *hiperenxebrismo* que probablemente se pueda interpretar como una mala traducción en el corpus de entrenamiento. El *hiperenxebrismo* o hipergalleguismo se puede definir como el fenómeno lingüístico que consiste en la manipulación de palabras gallegas con la intención de diferenciarlas o alejarlas de sus correspondientes en castellano, dando lugar a formas incorrectas que pretenden pasar por genuinas. Los tres revisores han marcado este error como grave.



ID	Source	Target	Category	Severity	Status	Description	TER
1443246	¿Sabías que la primera tarta de Santiago aparece documentada en 1577 durante una visita del Gobernador de Galicia D. Pedro de Porto durante una visita a la Universidad de Santiago, se llamaba entonces torta real?	¿Sabías que a primeira torta de Santiago aparece documentada en 1577 durante unha visita do escurecemento de Galicia D. Pedro de Porto durante unha visita á Universidade de Santiago, e que se chamaba entón torta real?	Terminology	Major	New	Gobernador	
1443276	Seas de la facultad de biología, ADE o ingeniería, ¡cubre el formulario de nuestra biografía para participar!	Sexas da facultade de bioloxía, ADE ou enxeñaría, cumprimenta o formulario da nosa biografía para participar!	Terminology	Major	New	"Cumprimentar" ten outro significado. Neste contexto unha opción sería o verbo "cubrir".	

Ilustración 56. Ejemplos de errores de terminología graves

Coincidimos con lo revisores en la gravedad del error y en la necesidad de resaltar este tipo de errores.

2.2.2. Errores graves de fluidez

A continuación, se presentan los errores de fluidez detectados, así como el nivel de acuerdo entre revisores al marcar la severidad de estos. También se indica aquí a qué tipo de tuit corresponden. Por un lado, tenemos un pronombre colocado en una posición incorrecta en una frase subordinada que pertenece a un hilo. Este error ha sido marcado como grave por dos revisores y como leve por uno.

Por otro, la ausencia de un acento en el verbo en un tuit formado por un párrafo de frases largas. En este caso, ha habido consenso en la gravedad marcada por los revisores.

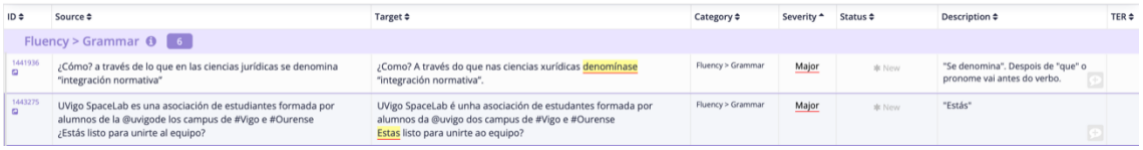


Ilustración 57. Ejemplos de errores graves de fluidez

Una posible interpretación de estos errores puede ser un mal uso de los pronombres y la presencia de errores ortográficos en el corpus específico.

2.2.3. Errores graves de precisión

A continuación, se presentan los errores de precisión detectados, así como el nivel de acuerdo entre revisores al marcar la severidad de estos. También se indica aquí a qué tipo de tuit corresponden.

Por un lado, encontramos una traducción incorrecta del tiempo verbal *reflexiornarmos* en lugar de *reflexionaremos*. Una posible interpretación de este error de traducción puede ser la presencia errores en el corpus de entrenamiento o por un mal entendimiento del tiempo verbal. Aquí no ha habido consenso entre los revisores: uno lo ha marcado como grave, otro no lo ha detectado y otro lo ha marcado como leve.

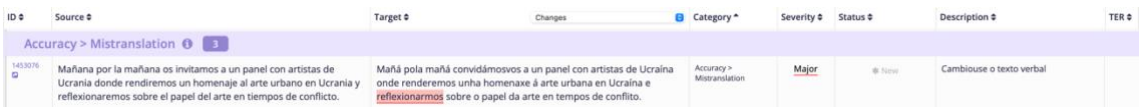


Ilustración 58. Ejemplo de error en un tiempo verbal

También encontramos elementos sin traducir, la abreviatura de siglo en gallego no se tradujo. La omisión de la sigla se puede deber a no disponer en el corpus de una lista de

abreviaturas y sus traducciones y a que aparece pegado al número con lo que se podría haber interpretado todo como una referencia. Tampoco se ha dado consenso en la severidad de este error.

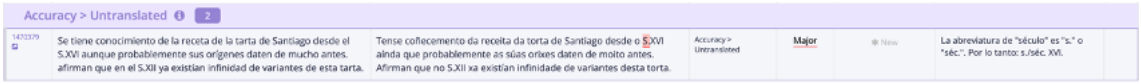


Ilustración 59. Ejemplo de elementos sin traducir

Por último, se ha encontrado una «sobretraducción» en el nombre del usuario de Twitter, probablemente debido a que al corpus específico orientado a que el motor sea capaz de traducir *hashtags*. En este error, sí ha habido consenso.

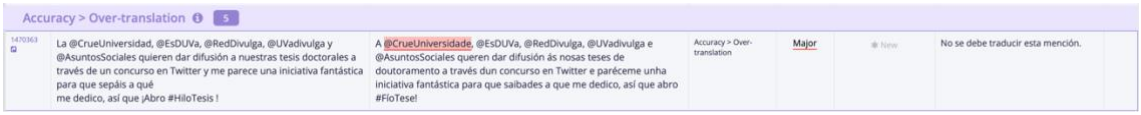


Ilustración 60. Ejemplos de sobretraducción

Dos de los errores se producen en hilos y el otro en una frase larga.

2.3. Comentarios generales de los revisores

Además de solicitar a los revisores que clasificaran los errores encontrados, también les dimos la posibilidad de que añadieran un comentario final sobre la calidad de la traducción. Podemos decir que en líneas generales los tres traductores coinciden en la buena calidad del resultado del motor de traducción automática. También destacan que muchos de los segmentos no necesitan ningún cambio y mencionan que en otros segmentos hay que realizar pequeños cambios para que la traducción sea totalmente correcta con respecto al castellano.

3. Evaluación mediante el principio de no inferioridad

Después del piloto, para ejecutar la evaluación de no inferioridad, se replica el estudio empírico con el diseño factorial de la prueba piloto y se realizan unos pequeños ajustes de acuerdo con las mejoras que los resultados del piloto sugirieron. Nos proponemos

obtener un mayor número de respuestas. Para ello decidimos elaborar cuestionarios más cortos. Lo que pretendemos es aumentar el número de informantes de cada tuit, y en global, de todos los cuestionarios. Por eso, en lugar de crear cuatro cuestionarios con un total de 20 ítems (tuits) de una muestra posible de 60 tuits, esta vez ampliamos los cuestionarios y diseñamos nueve cuestionarios con un total de 10 ítems cada uno de una muestra posible de 30 tuits, agrupados en 6 bloques diferentes de 5 tuits de diferente longitud cada uno. De nuevo, los tuits pueden ser tuits escritos originalmente en gallego extraídos de las mismas cuentas utilizadas para la creación del corpus, pero que no estén en los datos de entrenamiento; o tuits de la misma temática escritos en castellano y traducidos automáticamente con nuestro motor. Mantenemos los tuits según su longitud. De esta forma, distinguimos entre tuits formados por una frase corta de menos de 10 palabras, tuits formados por una frase larga de más de 10 palabras, tuits de un párrafo formado por frases cortas de menos de 10 palabras, tuits de un párrafo formado por frases largas formadas por más de 10 palabras e hilos. La categoría hilos se decide introducir tras los resultados del piloto con el objetivo de dotar de más contexto al participante.

De manera deliberada, se busca que la proporción de tuits originales y traducidos no fuera igual dentro de cada cuestionario para evitar que el participante pueda deducirlo por el simple formato de este. Así mismo, para garantizar la validez de las respuestas, cada cuestionario comparte ítems con los otros tal y como se puede ver en la siguiente tabla.

Cuestionario1	Cuestionario2	Cuestionario3	Cuestionario4	Cuestionario5	Cuestionario6	Cuestionario7	Cuestionario8	Cuestionario9
AFC2	AFC2	AFC2	Afío3	Afío3	Afío3	OPC3	OPC3	OPC3
APC2	APC2	APC2	OPC1	OPC1	OPC1	OFC1	OFC1	OFC1
APC1	APC1	APC1	AFC3	AFC3	AFC3	OFC3	OFC3	OFC3
AFL2	AFL2	AFL2	AFío2	AFío2	AFío2	Ofío3	Ofío3	Ofío3
OPL3	OPL3	OPL3	AFío1	AFío1	AFío1	OPL2	OPL2	OPL2
AFL3	APL3	APL1	AFL3	APL3	APL1	AFL3	APL3	APL1
APC3	OFL3	OFío1	APC3	OFL3	OFío1	APC3	OFL3	OFío1
AFL1	OPL1	APL2	AFL1	OPL1	APL2	AFL1	OPL1	APL2
OFío2	AFC1	OFC2	OFío2	AFC1	OFC2	OFío2	AFC1	OFC2
OFL2	OFL1	OPC2	OFL2	OFL1	OPC2	OFL2	OFL1	OPC2

Leyenda	
AFC	Tuit traducido automática compuesto por una frase corta
APC	Tuit traducido automática compuesto por un párrafo de frases cortas
AFL	Tuit traducido automática compuesto por una frase larga
AFío	Hilo traducido automáticamente
APL	Tuit traducido automática compuesto por un párrafo de frases largas
OFC	Tuit originalmente en gallego compuesto por una frase corta
OPC	Tuit originalmente en gallego compuesto por un párrafo de frases cortas
OFL	Tuit originalmente en gallego compuesto por una frase larga
OPL	Tuit originalmente en gallego compuesto por un párrafo de frases largas
OFío	Hilo escrito originalmente en gallego

Tabla 12. Distribución de los tipos de tuits en cada uno de los nueve cuestionarios

Como en la prueba piloto, se le presenta a los participantes el tuit aislado de uno en uno para evitar que anticiparan la existencia de algún patrón sistemático. Esta vez, el cuestionario, incluye información textual y paratextual: se muestra el tuit en su formato original tal y como se ve la aplicación (véase Ilustración 61 de la página siguiente).

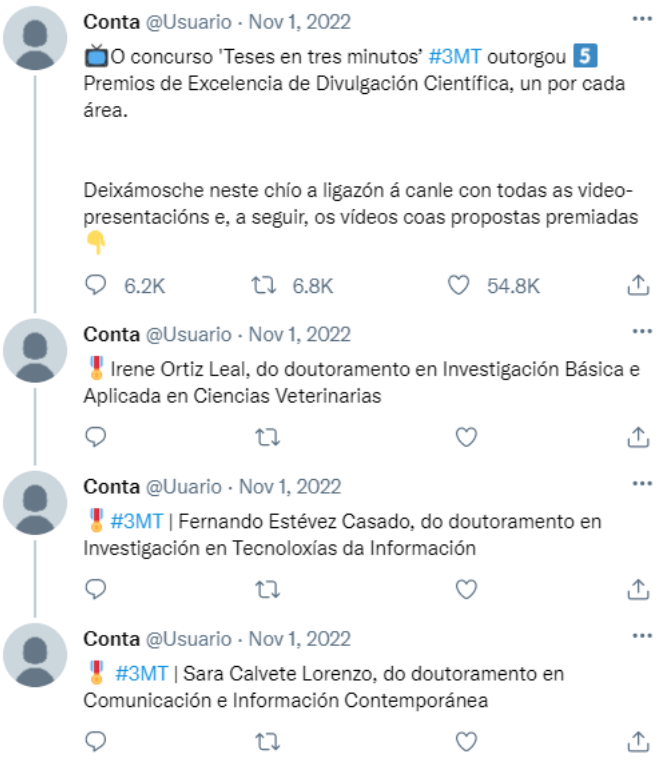


Ilustración 61. Ejemplo de cómo se presenta el tuit al participante. Tuit Ofío1.

Asimismo, se mantiene la pregunta de respuesta sí/no, pero se cambia la formulación de esta. Se pone el foco en la evaluación de la naturalidad sin pedir a los participantes que

juzguen si es un texto original o un texto traducido automáticamente. El objetivo es evitar mencionar «traducción automática» en la formulación para evitar que busquen deliberadamente segmentos traducidos automáticamente. La nueva pregunta es:

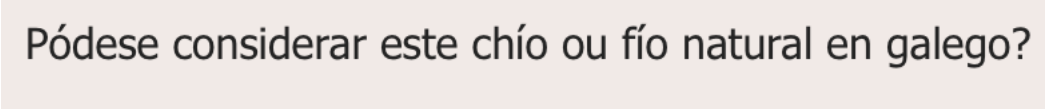


Ilustración 62. Pregunta principal del cuestionario de evaluación de la no inferioridad, tal y como aparece en el formulario

Debido a esta nueva formulación, se facilita al participante una explicación del concepto de naturalidad lingüística al inicio del cuestionario. La explicación escrita en gallego establece que se entiende la naturalidad como el uso habitual de la lengua, con expresiones comunes, sin errores gramaticales, con un estilo fluido y sin influencias de otras lenguas como el castellano. Esta explicación está incluida en la sección inicial de instrucciones e información del cuestionario (véase la Ilustración 63 de la página 154).

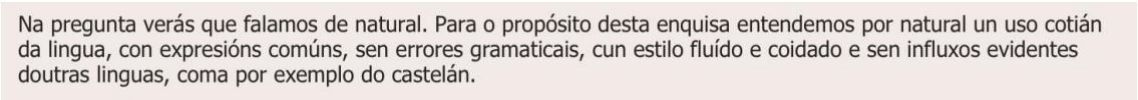


Ilustración 63. Explicación relativa a la naturalidad, tal y como aparece en el formulario

Los cuestionarios esta vez están dirigidos a hablantes gallegos de toda naturaleza, por lo que promocionamos la participación en el cuestionario dentro de organizaciones y asociaciones de todo tipo y no sólo lingüísticas.

Además de los 10 ítems o tuits, el cuestionario también se compone de cinco preguntas demoscópicas que decidimos introducir tras los resultados del piloto en cuanto a las expectativas lingüísticas de los participantes: edad, nivel de competencia lingüística en gallego, uso regular de gallego, nivel de uso de Twitter y nivel de formación recibida en gallego. En lo que respecta a la sección de tuits, al igual que en la prueba piloto, clasificamos los tuits en dos categorías amplias. La primera categoría es el origen: tuits escritos originalmente en gallego o tuit traducido automáticamente con el motor reentrenado. La segunda categoría tiene que ver con su longitud: frase corta, frase larga, párrafo de frases cortas, párrafo de frases largas e hilos.

3.1. Perfil de los participantes

En esta segunda evaluación, participaron un total de 228 personas que expresamente decidieron ceder los datos. Si bien, se excluyen del análisis algunos participantes por los siguientes motivos.

- Se eliminan 2 participantes por tener un tiempo de respuesta mayor que 70 minutos: ID: 321, 420, ya que el tiempo medio de respuesta se situaba entre 5-10 minutos. Se considera excluirlo por entender que tanto tiempo empleado en el cuestionario indica que no tienen las capacidades necesarias para completar el mismo.
- Se eliminan 4 participantes por no detectar ni un solo tuit como natural: ID: 204, 226, 311, 905. Decidimos excluir estos participantes por considerar que no entendieron el propósito del cuestionario, ya que desvirtúa todos los datos.
- Se eliminan 4 participantes por ser menores de edad: ID: 101,701,807,911. Decidimos excluir estos participantes al entender que todavía no tienen la madurez lingüística necesaria para realizar el cuestionario bien.

Por tanto, la base de datos de análisis consta de 218 participantes, tal y como podemos ver en el resumen de los datos descriptivos de los participantes de la Tabla 13.

	Total	**N**
	N=218	
Edad:		218
De 18 a 25	30 (13.8%)	
De 25 a 35	50 (22.9%)	
De 35 a 45	53 (24.3%)	
De 45 a 55	46 (21.1%)	
De 55 a 65	34 (15.6%)	
Maior de 65	5 (2.29%)	
Competencia lingüística en galego:		218
Nivel baixo como falante	7 (3.21%)	
Nivel medio como falante	34 (15.6%)	
Nivel alto como falante	135 (61.9%)	
Profesional (revisor, profesor de galego, tradutor, etc.)	42 (19.3%)	
Uso o galego:		218
Non, nunca	7 (3.21%)	
Maioritariamente o castelán	49 (22.5%)	

	Total	N
	N=218	
Na mesma proporción	39 (17.9%)	
Maioritariamente o galego	34 (15.6%)	
Si, sempre	89 (40.8%)	
Uso o galego en RRSS:		218
Non, nunca	18 (8.26%)	
Maioritariamente o castelán	49 (22.5%)	
Na mesma proporción	19 (8.72%)	
Maioritariamente o galego	42 (19.3%)	
Si, sempre	90 (41.3%)	
Formación reglada en galego:		218
Non	27 (12.4%)	
Si	191 (87.6%)	

Tabla 13. Resumen descriptivo de los datos demográficos

Esta vez el participante no tiene una edad media ya que obtuvimos una representación bastante heterogénea de todos los grupos de edad, excepto de los mayores de 65 años. Si bien, podemos decir que es un participante con un nivel alto de gallego, que emplea mayoritariamente el gallego en su día a día y en Twitter y que ha recibido formación reglada en gallego.

3.2. Procesamiento estadístico

Nuestra aproximación estadística, el principio de no inferioridad, sigue los mismos parámetros que la prueba piloto. Se emplea un modelo lineal mixto generalizado con una distribución binomial y una función *logit* para comparar la percepción de naturalidad de los tuits traducidos automáticamente y los tuits traducidos directamente. El modelo compara los tuits (ítem) y los participantes (ID), usando las características del tuit como factores fijos al disponer de información sobre los participantes y su uso y competencia lingüística. Esta vez, también comprobamos la percepción de naturalidad en función de la edad y del uso de la red social. Dado que los distintos tuits fueron evaluados por diferentes sujetos, el modelo considera al participante y al tuit como factores aleatorios y así nos aseguramos de que el modelo tenga en cuenta que el individuo evaluaba varios tuits y que había tuits que se repetían. Por último, realizamos una comparación detallada de la percepción de naturalidad de los tuits traducidos automáticamente y de los tuits escritos originalmente en gallego, que obtuvieron una tasa de identificación como tales

del 79,0 %, mediante una prueba unilateral de no inferioridad con un margen del 10 %. Nótese que el margen de inferioridad de la prueba piloto es del 20 % y esta vez quisimos ser aún más rigurosos y fijarla en un margen más exigente.

El modelo empleado es el siguiente:

$$
\begin{aligned}
\mathrm{N}_{ij[k]} &\sim \mathrm{Bin}\left(1, p_{ij[k]}\right) \\
\log\left[\frac{p_{ij[k]}}{1-p_{ij[k]}}\right] &= \mu + \alpha_i^{Id} + \alpha_{j[k]}^{Tw} + \beta \mathrm{O}_{\mathrm{Y}} + \gamma_j \mathrm{Type}_j + (\beta\gamma)_j\left(\mathrm{O}_{\mathrm{Y}} \times \mathrm{Type}_j\right) \\
\alpha_i^{Id} &\sim N\left(0, \sigma^2_{\alpha^{Id}}\right), \text{ for i = 1, ... ,I} \\
\alpha_{j[k]}^{Tw} &\sim N\left(0, \sigma^2_{\alpha^{Tw}}\right), \text{ for j= 1, ... ,J, k= 1, ... ,K\_j}
\end{aligned}
$$

Ecuación 2. Modelo de no inferioridad de la evaluación final

Donde:

p es la proporción de éxito (positivo)
α el término independiente del modelo
β un vector de coeficientes asociado a las variables explicativas
γ_j es el coeficiente asociado a la categoría j de la variable Tipo
$(\beta\gamma)_j$ es el coeficiente asociado a la categoria j de la interacción entre Original y Tipo
α_i es el efecto aleatorio asociado a los ítems, cuya variabilidad es $\sigma^2_{\alpha^{Tw}}$
α_j es el efecto aleatorio asociado a los sujetos, cuya variabilidad es $\sigma^2_{\alpha^{Id}}$

3.3. Resultados de la evaluación

En este apartado, describiremos primero la comparación descriptiva entre la percepción de naturalidad y las distintas variables (apartado 3.3.1 Análisis bivariado descriptivo), para, por último, realizar el análisis de no inferioridad (apartado 3.3.2 Análisis de no inferioridad).

Un punto crítico detectado en la prueba piloto para la viabilidad del instrumento es obtener las mismas respuestas en cada uno de los cuestionarios y, así, evitar que una distribución desproporcionada influencie en el análisis. Para evitar este punto crítico, nos aseguramos de crear cuestionarios cortos y un número equilibrado de respuestas. Por ello, se fueron controlando el número de respuestas de cada cuestionario para obtener un número similar de respuestas en los nueve cuestionarios sin grandes diferencias tal y como se ve en la Tabla 14.

Evaluación mediante el principio de no inferioridad

	Total
	N=218
Cuestionario:	
1	29 (13.3%)
2	24 (11.0%)
3	25 (11.5%)
4	23 (10.6%)
5	26 (11.9%)
6	21 (9.63%)
7	20 (9.17%)
8	30 (13.8%)
9	20 (9.17%)

Tabla 14. Distribución de respuestas en cada uno de los cuestionarios

Como podemos apreciar en la tabla de la página anterior, la distribución de respuestas obtenidas los porcentajes de respuesta se sitúan entre un 9 % y 13 %, por lo que podemos hablar de una distribución proporcionada y del éxito de la estrategia de control de respuestas.

3.3.1. Análisis bivariado descriptivo

El análisis bivariado descriptivo consiste en un análisis de los resultados cruzados de las variables observadas. Por tanto y de manera similar a la prueba piloto, realizamos una comparación de la percepción de naturalidad con las diferentes variables (número de cuestionario, grupo de edad, conocimiento del idioma, uso de la lengua, uso de la lengua en Twitter y formación) con el objetivo de detectar posibles anomalías que pudiesen llevar a fallos en el modelo. Solo encontramos diferencias significativas en dos variables: los distintos cuestionarios y el grupo de edad. Por tanto, no podemos concluir que la competencia en gallego, la formación reglada en gallego o el uso diario de gallego en la vida diaria y en las redes sociales pueda influir en los resultados. En el anexo IV, se puede consultar el análisis bivariado completo con todas las variables.

En cuanto a los cuestionarios, tal y como vemos en la Tabla 15, hay variaciones en la percepción de naturalidad dependiendo del cuestionario, aunque las tasas no distan mucho entre sí. En este caso, tal y como describimos en la evaluación de errores, se debe a una

mejor traducción del motor y a la ausencia de errores graves. Por ejemplo, el cuestionario 6 se compone del hilo marcado con un error grave de terminología o el cuestionario 7 del tuit de frase larga marcado con un error grave de fluidez.

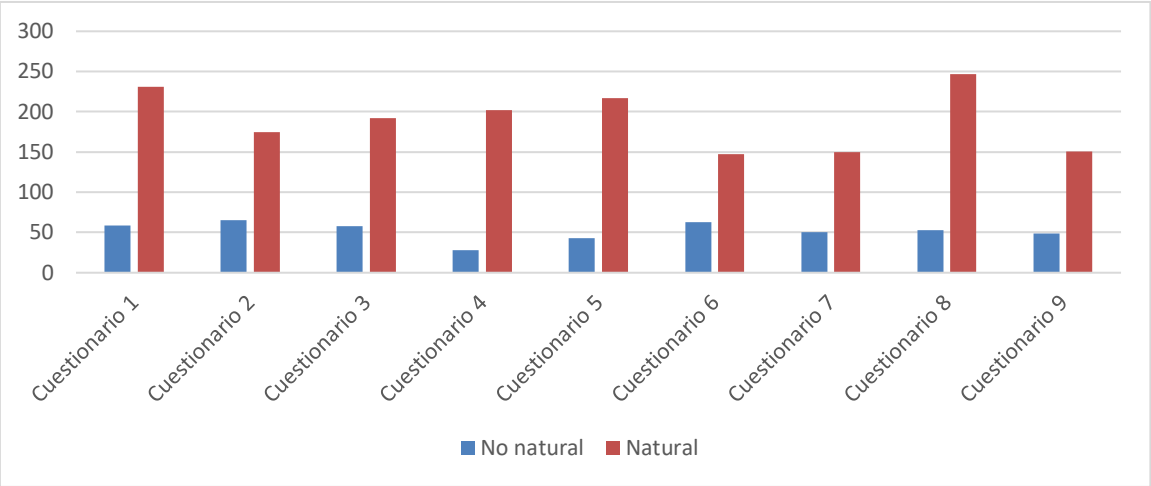


Tabla 15. Análisis bivariado descriptivo según cuestionario

En cuanto a la variable edad, en la Tabla 16, vemos que los grupos de entre 18 y 25 años (85,3 %) y de entre 35 y 45 (83 %) son los que más aceptan el tuit traducido automáticamente como original.

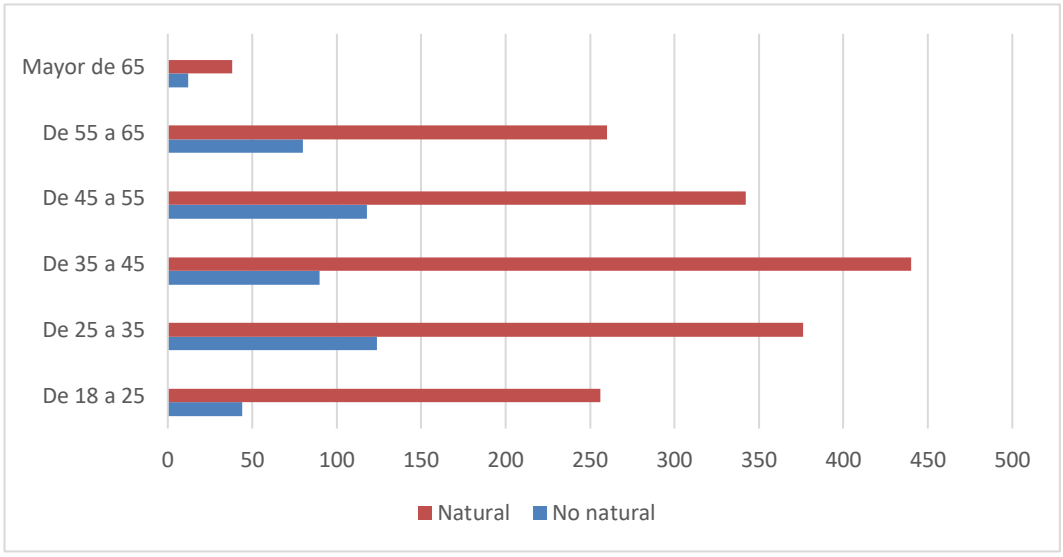


Tabla 16. Análisis bivariado descriptivo según grupos de edad

Por el contrario, si ahora nos fijamos en los participantes de entre 45 y 55 años, podemos apreciar que son los más exigentes, a pesar de que son el grupo con menor formación en gallego, si lo comparamos con los anteriores.

3.3.2. Análisis de no inferioridad

El modelo de inferioridad empleado asume que los distintos cuestionarios comparten tuits, por lo que algunos tuits han sido evaluados por más de una persona. Adicionalmente, cada tuit se corresponde con una categoría única de las posibles categorías. Dado que queremos ser lo más exigentes posible con nuestro análisis, empleamos el modelo que tiene en cuenta estas particularidades.

Al realizar un análisis global de no inferioridad, podemos declarar la no inferioridad global, ya que obtuvimos un *p*-valor de 0,024, lo que valida nuestra hipótesis.

```
Analysis of Deviance Table (Type III Wald chisquare tests)

Response: Naturalidad2
                     Chisq Df Pr(>Chisq)
(Intercept)         32.3833  1  1.266e-08 ***
Original             0.4464  1   0.504048
TipoTweet2          13.5939  4   0.008710 **
Original:TipoTweet2 14.5260  4   0.005792 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  contrast odds.ratio    SE  df  null z.ratio p.value
1 Non / Si      1.266 0.231 Inf 0.884   1.972   0.024
```

Ilustración 64. Cálculo global de no inferioridad siguiente el modelo diseñado

No obstante, para tener una idea clara de los puntos fuertes y débiles del motor reentrenado, se replica este análisis de no inferioridad teniendo en cuenta la variable tipo de tuit según su longitud. Así, si observamos la longitud de los tuits, encontramos un compartimiento diferente y dos grupos diferenciados. Por un lado, según nuestros participantes, los tuits formados por un párrafo de frases cortas o hilos son las categorías percibidas como menos naturales. Estas dos categorías son las que en la Ilustración 65 obtienen un valor de probabilidad de aceptación por debajo de los tuits originales (columna *prob*). Por el contrario, los tuits formados por una frase larga o un párrafo de frases largas fueron percibidos como los más naturales (valores 0,946 y 0,900, respectivamente), superando la tasa obtenida por la misma categoría de tuits originales

con un margen considerable (valores 0,888 y 0,745, respectivamente) y que se podía prever tras analizar los resultados del piloto.

	Original	TipoTweet2	prob	SE	df	asymp.LCL	asymp.UCL
1	Non	FC	0.846	0.039	Inf	0.754	0.908
2	Si	FC	0.809	0.047	Inf	0.701	0.885
3	Non	fío	0.897	0.031	Inf	0.820	0.943
4	Si	fío	0.915	0.025	Inf	0.851	0.953
5	Non	FL	0.946	0.018	Inf	0.899	0.972
6	Si	FL	0.888	0.031	Inf	0.812	0.936
7	Non	PC	0.815	0.045	Inf	0.709	0.888
8	Si	PC	0.907	0.027	Inf	0.838	0.948
9	Non	PL	0.900	0.029	Inf	0.828	0.944
10	Si	PL	0.745	0.055	Inf	0.624	0.837

Ilustración 65. Cálculo de probabilidad de aceptación de los tuits según su longitud

Por su parte, los tuits de una frase corta (valor 0,846) se percibieron como más naturales que los párrafos de frases cortas (valor 0,815) y su tasa de aceptación es superior a la obtenida en la prueba piloto, lo que podría ser un resultado directo del reentrenamiento del motor.

Para poder establecer la no inferioridad, el *p*-valor de la Ilustración 66 debe ser inferior a 0,025. Por lo tanto y tal y como podemos apreciar en los valores marcados en negrita en la columna *p.value*, únicamente podemos declarar la no inferioridad en los tuits de frases largas y párrafos compuestos por frases largas. En el resto de las categorías, no encontramos diferencias significativas, por lo que no podemos declarar la no inferioridad

	contrast	TipoTweet2	odds.ratio	SE	df	null	z.ratio	p.value
1	Non / Si	FC	1.297	0.505	Inf	0.876	1.007	0.157
2	Non / Si	fío	0.810	0.341	Inf	0.891	-0.226	0.589
3	Non / Si	FL	2.222	0.935	Inf	0.887	2.181	**0.015**
4	Non / Si	PC	0.453	0.183	Inf	0.890	-1.669	0.952
5	Non / Si	PL	3.079	1.195	Inf	0.866	3.268	**0.001**

Ilustración 66. Análisis de no inferioridad según longitud del tuit

Para entender mejor esta tasa que permite observar la naturalidad, miramos con más detalle las traducciones del motor y contrastamos con los errores de la evaluación humana (véase Tabla 17 de la página 162).

En cuanto a los hilos, detectamos que, aunque se preprocesaron como una sección completa, nuestro motor no tiene en cuenta el contexto, por lo que la cohesión se ve afectada. Además, como vimos en el apartado 2, en estos tuits se concentran una mayor cantidad de errores tanto graves como leves.

Una posible interpretación para la falta de naturalidad en los tuits formados por párrafos de frases cortas puede ser su estilo crítico y telegráfico, lo que pone en riesgo la fluidez. Si bien, los revisores detectaron aquí un error grave y varios leves. Por el contrario, el éxito en los tuits de frase y párrafo largos se pueda deber a que, al no haber restricciones de caracteres, la fluidez se veía favorecida y los errores leves encontrados podrían haber sido más tolerados.

Longitud del tuit	Errores leves	Errores graves
Hilos	13	4
Frase corta		
Frase larga	1	1
Párrafo corto	4	1
Párrafo largo	6	1

Tabla 17. Comparación de gravedad de errores de todo tipo con la longitud del tuit

Por último, en lo relativo a los tuits formados por una frase corta, no se identifican errores en ellos. Sin embargo, sí se constata que se trata de una traducción muy literal al castellano y sin expresiones únicas con las que distinguirse del castellano. Debido a la relación diglósica del castellano y gallego, esto se podría interpretar como menos natural en gallego.

3.4. Variaciones del modelo de no inferioridad

Con el objetivo de validar la consistencia de los resultados obtenidos y explorar los límites de nuestra formulación, se modificó el modelo de análisis para eliminar el efecto que puede tener en el análisis de los resultados los efectos aleatorios, es decir, la consideración de que una persona evalúa más de un tuit y que los tuits se repiten en los cuestionarios. Con estas modificamos, pretendemos estresar el modelo para demostrar que nuestra fórmula tiene en cuenta la naturaleza del diseño para interpretar las respuestas consistentemente, y además, determinar los límites de esta de cara a su replicabilidad.

Estos ejemplos de modificaciones pueden proporcionar a los investigadores opciones adicionales para interpretar su base de datos, particularmente si no disponen de una muestra muy grande de respuestas o de tuits. El modelo se puede adaptar a las necesidades particulares de la combinación lingüística o de la población de muestra.

En una primera modificación de nuestra fórmula, hicimos que el modelo siguiera interpretando que un mismo participante evaluaba varios tuits, pero tratamos cada tuit como si fuese único y no estuviese repetido en los distintos cuestionarios. Esta es la fórmula del modelo.

$$\begin{aligned} \mathrm{N}_{ij} &\sim \mathrm{Bin}\left(1, p_{ij}\right) \\ \log\left[\frac{p_{ij}}{1-p_{ij}}\right] &= \mu + \alpha_i^{Id} + \beta \mathrm{O_Y} + \gamma_j \mathrm{Type}_j + (\beta\gamma)_j\left(\mathrm{O_Y} \times \mathrm{Type}_j\right) \\ \alpha_i^{Id} &\sim N\left(0, \sigma^2_{\alpha^{Id}}\right), \text{ for i = 1, ... ,I} \end{aligned}$$

Ecuación 3. Variación del modelo sin el efecto aleatorio tipo de tuit

Como vemos en la siguiente figura, el *p*-valor global de no inferioridad es igual al obtenido en el modelo anterior, lo que nos permite declarar aquí también la no inferioridad global.

```
Analysis of Deviance Table (Type III Wald chisquare tests)

 Response: Naturalidad2
                      Chisq Df Pr(>Chisq)
 (Intercept)         54.6570  1  1.435e-13 ***
 Original             0.6205  1     0.4309
 TipoTweet2          25.5071  4  3.977e-05 ***
 Original:TipoTweet2 28.0446  4  1.222e-05 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  contrast odds.ratio    SE  df  null z.ratio p.value
 1 Non / Si       1.26 0.167 Inf 0.884   2.686   0.004
```

Ilustración 67. Cálculo global de no inferioridad siguiente la primera modificación del modelo.

En esta modificación del modelo y tal y como se puede ver en la columna *p.value* de la Ilustración 68, los *p*-valores del análisis son inferiores que en el modelo anterior (véase Ilustración 66 de la página 161). Aun así, la no inferioridad se consigue en las mismas categorías anteriores.

```
  contrast TipoTweet2 odds.ratio    SE  df  null z.ratio p.value
1 Non / Si         FC      1.239 0.337 Inf 0.876   1.273   0.102
2 Non / Si        fío      0.815 0.257 Inf 0.890  -0.281   0.611
3 Non / Si         FL      2.233 0.700 Inf 0.887   2.945   0.002
4 Non / Si         PC      0.458 0.132 Inf 0.889  -2.297   0.989
5 Non / Si         PL      3.081 0.827 Inf 0.865   4.731   0.000
```

Ilustración 68. Análisis de no inferioridad según longitud del tuit con la primera modificación del modelo

En segundo lugar, fuimos un paso más allá y eliminamos el efecto del participante en la fórmula. De esta manera, evaluamos las respuestas usando un modelo que interpreta todas las respuestas como si fuesen de personas diferentes y cada uno de los tuits de los cuestionarios como único, tal y como se ve en la siguiente fórmula:

$$\begin{aligned} \mathrm{N}_{ij} &\sim \mathrm{Bin}\left(1, p_{ij}\right) \\ \log\left[\frac{p_{ij}}{1-p_{ij}}\right] &= \mu + \beta \mathrm{O}_{\mathrm{Y}} + \gamma_j \mathrm{Type}_j + (\beta\gamma)_j\left(\mathrm{O}_{\mathrm{Y}} \times \mathrm{Type}_j\right) \end{aligned}$$

Ecuación 4. Variación del modelo sin el efecto aleatorio tipo de tuit y del participante

En este caso, a se pueden apreciar más diferencias con los resultados anteriores. Por un lado, el *p*-valor global (columna *p.value* de la Ilustración 69) es inferior a los dos primeros, lo que permite declarar la no inferioridad global con más holgura.

```
Analysis of Deviance Table (Type III tests)

Response: Naturalidad2
                   LR Chisq Df Pr(>Chisq)
Original             1.3793  1  0.2402137
TipoTweet2          22.1636  4  0.0001859 ***
Original:TipoTweet2 19.4415  4  0.0006435 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  contrast odds.ratio    SE  df  null z.ratio p.value
1 Non / Si      1.185 0.129 Inf 0.872    2.83   0.002
```

Ilustración 69. Cálculo global de no inferioridad siguiente la segunda modificación del modelo.

Por otro lado, al analizar cada categoría de tuits, vemos que los *p*-valores de la columna *p.value* en la Ilustración 70 son lo suficientemente bajos como para también declarar la no inferioridad en los tuits compuestos por frases cortas.

```
   contrast TipoTweet2 odds.ratio    SE  df  null z.ratio p.value
 1 Non / Si         FC      1.293 0.283 Inf 0.860   1.862   0.031
 2 Non / Si        fío      0.884 0.229 Inf 0.881   0.017   0.493
 3 Non / Si         FL      1.742 0.474 Inf 0.878   2.516   0.006
 4 Non / Si         PC      0.573 0.134 Inf 0.878  -1.819   0.966
 5 Non / Si         PL      2.044 0.458 Inf 0.851   3.916   0.000
```

Ilustración 70. Análisis de no inferioridad según longitud del tuit con la segunda modificación del modelo

De esta forma, con este modelo, tres de las cinco categorías serían no inferiores: tuits formados por frases cortas, tuits formados por frases largas y tuits formados por párrafos de frases largas.

3.5. Consideraciones finales

En este capítulo dedicado a la triangulación de métodos de evaluación diferentes, tanto cuantitativos como cualitativos, se ha explorado en detalle la calidad de las traducciones generadas por la segunda versión del motor desarrollado y se han analizado las percepciones de los usuarios en relación con la naturalidad de dichas traducciones. A través del estudio realizado, se han obtenido hallazgos intrigantes que arrojan luz sobre la percepción de la naturalidad de los usuarios finales en las traducciones automáticas.

Se constata una concentración elevada de errores graves en los hilos y en los párrafos cortos. Este resultado podría explicar la baja naturalidad percibida en el estudio de no inferioridad. Esta presencia de errores puede indicar que los usuarios son más sensibles a los errores graves que afectan a la comprensión y fluidez del texto. De esta forma, vemos que el modelo detecta no sólo la naturalidad, si no también aquellas categorías de errores en las que no es necesario acceder al original (fluidez y no traducción de léxico). Sin estos errores, la percepción general de las traducciones automáticas de estas categorías de tuits sería superior. Valida, por tanto, la metodología de no inferioridad para detectar segmentos que necesiten de una posedición antes de su publicación directa.

Por otro lado, se encontró que los errores leves estaban distribuidos de manera indiscriminada en todos los tipos de tuits. Sin embargo, se observó que la gravedad de estos errores no afecta significativamente a las percepciones de los usuarios en cuanto a la naturalidad de las traducciones. Esto sugiere que los usuarios pueden ser más tolerantes con los errores leves, siempre y cuando no comprometan la comprensión global del mensaje.

Estos hallazgos tienen implicaciones importantes para la mejora continua del motor de traducción automática. El hecho de que los errores graves se concentren en hilos y párrafos cortos indica la necesidad de desarrollar estrategias específicas para abordar estos casos y mejorar la coherencia y fluidez de estos.

La triple evaluación realizada ha permitido brindar información valiosa sobre la percepción de los usuarios en relación con los errores de las traducciones. Estos hallazgos respaldan la importancia de realizar una evaluación rigurosa y continua de los sistemas de traducción automática y proporcionan una base sólida para futuras mejoras en el motor desarrollado. El análisis de los errores identificados y su impacto en la percepción de los usuarios permitió obtener una visión clara sobre la calidad y naturalidad del motor desarrollado y, lo más importante, la consolidación del método de no inferioridad para evaluar la naturalidad de la traducción automática. Podemos afirmar, por tanto, que hemos cumplido con nuestro doble objetivo de crear un motor de TAN para gallego y las redes sociales y proponer una metodología de evaluación de la TAN novedosa. Los resultados obtenidos validan la hipótesis de partida de que se podía alcanzar una calidad alta si se controlaban los parámetros de entrenamiento mediante el corpus específico de Twitter. Asimismo, refutamos también la hipótesis de la aproximación a la TAN debe realizarse desde una perspectiva diferente a como se venía haciendo hasta el momento. Con nuestra investigación, abrimos las puertas a otras lenguas minorizadas en una situación similar al gallego para que repliquen nuestro proceso de entrenamiento y evaluación y alcancen objetivos similares. Para ello, pueden utilizar los modelos estadísticos propuestos y ajustar el margen de no inferioridad dependiendo del tamaño de la muestra y del número de participantes.

Por último, también consideremos que nuestro modelo de evaluación se puede utilizar en cualquier combinación lingüística dentro de un entorno profesionalizado de la traducción para detectar cuando un segmento se debe poseditar.

CAPÍTULO XI: Conclusiones

Esta investigación surgió para dar respuesta a través de las tecnologías de la traducción a necesidades del gallego en tanto que lengua con pocos recursos digitales, ya que son estos últimos el requisito imprescindible para obtener el mayor beneficio posible de las tecnologías emergentes de los últimos años. Así, nos propusimos alcanzar una serie de objetivos relacionados con la creación, uso y evaluación de un motor de traducción automática neuronal (TAN) al gallego como ejemplo de lengua con pocos recursos y podemos afirmar que se han alcanzado con creces. A continuación, a partir de dichos objetivos, revisaremos los resultados obtenidos y si estos permiten refutar las hipótesis de partida de esta tesis.

1. Resultados

Uno de los primeros objetivos consistía en profundizar en las prácticas comunes de la industria a través de un estudio de campo. Fruto de este estudio, pudimos partir de la base de que necesitábamos, por un lado, un corpus de como mínimo de entre 500 000 y 1 000 000 de pares de segmentos y, por otro, diseñar una estrategia de evaluación acorde a las lenguas y género textual seleccionados.

En relación con el objetivo de generar un corpus suficiente en calidad y cantidad, pudimos combinar el corpus genérico *Paracrawl* con más de un millón de pares con un corpus específico de Twitter de 262 785 oraciones únicas bilingües y 5 448 375 palabras, que descargamos, limpiamos y compilamos. Así, pudimos demostrar que el uso de datos específicos monolingües de Twitter y la técnica de traducción inversa o *backtranslation* son estrategias efectivas en el desarrollo del motor para el ámbito de las redes sociales en combinaciones de idiomas con recursos limitados. Mediante estas dos estrategias, se ha logrado fijar un número mínimo de segmentos necesarios para el entrenamiento de un motor neuronal en la combinación de trabajo y tipo de texto elegido, ampliar los datos generales disponibles y, con ello, mejorar el rendimiento del motor tanto en términos de evaluación automática como humana. Es importante destacar que los resultados no permiten establecer con seguridad el tamaño óptimo de un corpus para el entrenamiento u optimización de un motor, pero sí permiten presumir que la proximidad y la complejidad

de los idiomas empleados sea determinante. Dado que el castellano y el gallego son lenguas muy cercanas, se han obtenido resultados prometedores incluso con un corpus específico de Twitter reducido. Asimismo, se ha demostrado que Twitter es una fuente de datos monolingües valiosa para la extracción de datos lingüísticos, y no solo para usos comunes como el análisis de sentimientos (Zimbra *et al.*, 2018). Su utilización en el desarrollo del motor ha sido un factor clave para la mejora del rendimiento y la adaptación a la especificidad del lenguaje utilizado en las redes sociales. A través de la consecución de este objetivo se confirma la hipótesis inicial según la cual, con la tecnología actual y los recursos lingüísticos producidos a través de las redes sociales, es posible generar un conjunto de datos que, en calidad y en volumen, sean suficientes para crear un motor de TAN con una calidad igual o superior a la de un motor de lenguas dotadas de más recursos.

En cuanto al objetivo relativo a la creación de un motor de alta calidad, cabe mencionar que, gracias a la metodología de evaluación aplicada, hemos obtenido información clara y significativa sobre las debilidades y fortalezas. La evaluación triple realizada, basada en métodos automáticos como BLEU, en la evaluación humana siguiendo el marco MQM-DQF y en el principio de no inferioridad, demostró que nuestro motor es capaz de traducir contenidos propios de las redes sociales con una gran solvencia. Si bien, las estimaciones basadas en el modelo nos han mostrado dónde se abre una puerta para una futura mejora del rendimiento del motor o dónde debemos realizar una posedición de la traducción automática en bruto: en los tuits con frases cortas, especialmente cuando se presentan en un párrafo o en hilos, ya que requieren más trabajo para sonar naturales. No obstante, queremos remarcar especialmente que nuestro motor no fue inferior a los textos originalmente escritos en frases y párrafos largos, es decir, aquellos que requieren una mayor cohesión lingüística. Este dato resulta llamativamente prometedor si tenemos en cuenta que no se realizó ninguna posedición. Esto nos lleva a concluir que el uso de la TAN como impulso para la comunicación bilingüe en una red social como Twitter es prometedor, ya que se puede brindar acceso a lectores finales sin experiencia en traducción automática (Guerberof y Toral, 2022), lo que demuestra la viabilidad de una TAN para las redes sociales. Por último, podemos refutar nuestra hipótesis de que nos encontraríamos con diferentes niveles de calidad de la TAN en función de la longitud del segmento que se va a traducir.

En cuanto a la evaluación, hemos podido constatar cómo, con una TAN de alta calidad, los sistemas de evaluación basados en la comparación con traducción humana son insuficientes. Por eso, elaboramos una propuesta innovadora basada en la percepción de los hablantes en la lengua de llegada y en el principio estadístico de no inferioridad. Al usar este principio, pretendíamos determinar si los textos traducidos automáticamente desde el castellano al gallego eran menos naturales o no que los textos creados originalmente en gallego. Gracias al análisis de la prueba piloto y a la evaluación final, podemos concluir que este método es efectivo para recopilar las percepciones de los usuarios finales sobre la TAN. Aunque las evaluaciones de no inferioridad son comúnmente utilizadas en pruebas farmacológicas, hemos demostrado que también pueden ser utilizadas en las evaluaciones de traducción automática. Al comparar los tuits traducidos por el motor con los tuits originalmente escritos en gallego, nuestro método se centró en identificar diferencias de percepción. El método puso de manifiesto que los textos originales tampoco se perciben siempre como 100 % naturales. La comprobación de este fenómeno contribuye a poner en duda los métodos basados en la comparación de traducciones automáticas con otras traducciones humanas, ya que cualquier texto puede ser percibido como no natural, ya sea traducción o texto original. Gracias a la consecución de este objetivo, podemos, por tanto, afirmar que se cumple la hipótesis de que la TAN para las redes sociales en un contexto de lenguas con pocos recursos necesitaría de una nueva metodología de evaluación. Así, proponemos el enfoque de no inferioridad como un método apropiado para evaluar la naturalidad de la traducción automática sin comparar las traducciones entre sí.

A pesar de que no entraba dentro de nuestros objetivos, el estudio de no inferioridad también reveló que existen diferencias en la aceptación de la naturalidad entre los grupos de edad. Los participantes más jóvenes mostraron una mayor aceptación que los participantes de mediana o gran edad. Se abre, por tanto, aquí una ventana de estudio en el futuro centrada en entender por qué los usuarios más jóvenes son menos reacios a aceptar cambios en el idioma que los usuarios más mayores y ver si estas conclusiones pueden evitar la pérdida de hablantes en este grupo de edad mencionada en la introducción de esta tesis. Sin embargo, cabe resaltar también que esta diferencia detectada pudiera verse exacerbada por el género, es decir, tuits, y hubiese contado con índices menores y menos perceptibles en un género más tradicional como en textos más largos y cohesivos.

2. Interés y aplicación de los resultados

En primer lugar, queremos destacar la importancia de llevar a cabo investigaciones en TAN para lenguajes con recursos reducidos. Nuestro proceso de entrenamiento y, en particular, la utilización de la técnica del *backtranslation* se puede tomar como referencia en otras combinaciones de idiomas similares a la nuestra para crear un motor de TAN de alta calidad. Hemos demostrado que en un género libre como tuits esta técnica puede funcionar, lo que la convierte en un método útil entre muchos géneros en la combinación GL-ES, y en un modo válido de aumentar los datos paralelos para la lengua con pocos recursos.

Con nuestra investigación, esperamos contribuir a que el gallego llegue a la población más joven a través de las redes sociales y a reducir la pérdida de hablantes en los sectores más jóvenes de la población que se ha venido produciendo en las últimas décadas. También esperamos que haber puesto a disposición de los usuarios finales el motor de traducción a través de un entorno amigable como el del proyecto MuTNMT y el modelo de lengua en la plataforma Github ayude a contribuir con la investigación de recursos de procesamiento natural del gallego y amplíe los trabajos en esta temática. En un futuro, haremos también accesible el modelo marian del motor y el corpus específico a través del repositorio de la UAB (Dipòsit Digital de Documents).

En cuanto a la transferencia de nuestro procedimiento de evaluación, consideramos que este método puede ser utilizado en diversas situaciones dentro del mundo académico y profesional de la traducción automática. Por un lado, en el caso de idiomas con pocos datos, sería un método válido cuando el hablante de ese idioma es el usuario final del motor de traducción automática. Esto es especialmente común en lenguajes con recursos reducidos, donde la traducción automática es otro recurso para promover el uso del lenguaje y la presencia en Internet a través del uso de motores adecuados para cada propósito (Van Edgom, 2019). Por otro lado, nuestro método también puede ser empleado en la industria de la traducción para ayudar a los proveedores de servicios lingüísticos a determinar si una traducción automática con alta fluidez y precisión requiere posedición o puede ser publicada directamente en contextos multilingües y publicaciones instantáneas (redes sociales, comunicados de prensa, entradas de blog, artículos de ayuda, etc.). Creemos que debe combinarse con las metodologías actuales basadas en precisión y fluidez para obtener una visión holística de la calidad del motor. Finalmente, también

podría ser utilizado en investigación académica, ya que comparar la no inferioridad de textos de TA con la de los mismos textos poseditados podría permitir establecer de manera explícita el valor añadido de la posedición, no solo segmento por segmento sino a un nivel textual, teniendo en cuenta la coherencia y la cohesión textuales.

3. Perspectivas de trabajo en el futuro

El presente estudio ha sentado las bases para investigaciones futuras en TAN para el idioma gallego, usando estrategias de ampliación de datos y un enfoque novedoso de evaluación. No obstante, existen diversas oportunidades de investigación y mejoras que pueden ser exploradas en el futuro.

Una de las áreas de investigación prometedoras es la aplicación de modelos de lengua más grandes, como los *Large Language Models* (LLM), en el desarrollo de motores de TAN. Los LLM han demostrado un rendimiento sobresaliente en una amplia gama de tareas del procesamiento del lenguaje natural (PLN). Estos modelos, entrenados en enormes cantidades de datos, muchas veces monolingües, capturan de manera efectiva las complejidades y sutilezas del lenguaje, lo que puede conducir a una mejora significativa en la calidad de la traducción automática.

En este sentido, los resultados de esta tesis establecen una base de referencia para investigaciones futuras sobre TAN y LLM, que podrían centrarse en el estudio de cómo la utilización de modelos de lengua más grandes puede influir en la calidad de la traducción automática para idiomas con recursos reducidos. En concreto, se podrían llevar a cabo investigaciones para evaluar y comparar el rendimiento de diferentes LLM en la traducción automática del gallego. Esto permitiría identificar los modelos más adecuados y optimizar su configuración para obtener los mejores resultados en términos de calidad de traducción.

Otra área de investigación interesante sería la expansión del enfoque propuesto de evaluación de la traducción automática a los LLM. Dado que estos modelos están demostrando la capacidad de poder hacer traducciones aparentemente condicionadas por parámetros expresados en el *prompt* (Sánchez-Gijón y Palenzuela-Badiola, 2023), se podría explorar la posibilidad de adaptar el principio estadístico de no inferioridad para evaluar la naturalidad y la calidad de las traducciones generadas por los LLM en

comparación con textos escritos originalmente en el idioma de destino. Asimismo, también cabría explorar los resultados que podría arrojar un LLM optimizado con un corpus monolingüe en gallego o paralelo como el creado en esta tesis para ver si la presencia de datos concretos en la lengua con pocos recursos mejora el rendimiento de un LLM entrenado mayoritariamente con datos en otras lenguas.

En resumen, el futuro trabajo de investigación podría enfocarse en la aplicación y el estudio de los LLM en el desarrollo de motores de TAN para el gallego, así como en la adaptación del enfoque de evaluación propuesto a estos modelos más grandes tanto para fines académicos como en un entorno industrial o profesional. Estas investigaciones podrían contribuir a mejorar aún más la calidad de la traducción automática y promover el uso del gallego en el contexto de las redes sociales y la comunicación en línea.

Bibliografía

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Althunian, T. A., de Boer, A., Groenwold, R. y Olaf H. K. (2017). Defining the noninferiority margin and analysing noninferiority: an overview. En *British Journal of Clinical Pharmacology*, 83(8), 1636-1642.

Asamblea General de la ONU. (1948). *Declaración Universal de los Derechos Humanos* (217 [III] A). Paris.

Asamblea General de la ONU. (1966). *Pacto Internacional de Derechos Económicos, Sociales y Culturales.* Adoptado y abierto a la firma, ratificación y adhesión por la Asamblea General en su resolución 2200 A (XXI), de 16 de diciembre de 1966, Naciones Unidas, Serie de Tratados, vol. 993, p. 3, disponible en esta dirección: https://www.refworld.org.es/docid/4c0f50bc2.html [Fecha de consulta: 2 de septiembre de 2023].

Asamblea General de la ONU. (1966). *Pacto Internacional de Derechos Civiles y Políticos*. Adoptado y abierto a la firma, ratificación y adhesión por la Asamblea General en su resolución 2200 A (XXI), de 16 de diciembre de 1966, Naciones Unidas, Serie de Tratados, vol. 999, p. 171, disponible en esta dirección: https://www.refworld.org.es/docid/5c92b8584.html [Fecha de consulta: 2 de septiembre de 2023].

Asamblea General de la ONU. (2002). *Resolución A/RES/56/262 .aprobada por la Asamblea General*.

Asamblea parlamentaria del Consejo de Europa. (1981). *Recommendation 928 on Educational and cultural problems of minority languages and dialects in Europe*.

Bahdanau, D., Kyunghyun, C. y Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. En Bengio, Y. y LeCun, Y. (Eds.). *Proceedings of the 3rd International Conference on Learning Representations*, ICLR 2015. San Diego.

Banerjee S. y Lavie A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. En *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 65–72.

Bañón M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D. y Zaragoza, J.. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4555–4567. Online. Association for Computational Linguistics.

Bentivogli L., Cettolo M., FedericoM. y Federmann C. (2018). Machine translation human evaluation: An investigation of evaluation based on Post-editing and its relation with Direct Assessment. En *Proceedings of the International Workshop on Spoken Language Translation*, 62–69. Brujas, Bélgica.

Bhardwaj, S., Hermelo, D. A., Langlais, P., Bernier-Colborne, G., Goutte, C. y Simard, M. (2020). Human or Neural Translation? En *Proceedings of the 28th International Conference on Computational Linguistics*, 6553–6564. Barcelona, España.

Blatz, J., Fitzgerald, E., Foster, G., Gandraburn, S., Goutte, C., Kulesza, A., Sanchis, A. y Ueffing, N. (2004). Confidence estimation for machine translation. En *Proceedings of the 20th International Conference on Computational Linguistics (COLING*). Ginebra, Suiza.

Bosca, A., Dini, L., Kouylekov, M. y Trevisan, M. (2012). Linguagrid: a network of Linguistic and Semantic Services for the Italian Language. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3304–3307. Estambul, Turquía.

Casares Berg, H. y Monteagudo Romero, H. (2021). La juventud gallega y las pantallas. Una aproximación sociolingüística. En *Quaderns del CAC* 47, vol. XXIV, 39-49.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley J., & Way, A. (2017). Is neural machine translation the new state of the art? En *The Prague Bulletin of Mathematical Linguistics*, 108, 109-120.

Casucuberta Nolla, F. y Peris Abril, Á. (2017). Traducción automática neuronal. En *Revista Tradumàtica. Tecnologies de la Traducció*, 15, 66-74.

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. En *Natural Language Engineering*, 26(2), 137-161. Cambridge University Press. https://doi.org/10.1017/S1351324919000469

Choudhury,M., Counts, S. y Gamon, M. (2010). Not all moods are created equal! Exploring human emotional states in social media. En *Proceedings of the 2nd ACM SIGMM workshop on Social media*.

Chowdhury, K.D., Hasanuzzaman, M., Liu, Q. (2018). Multimodal neural machine translation for low-resource language pairs using synthetic data. En *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, 33-42. Melbourne, Australia, 19 de julio de 2018.

Cieri, C., DiPersio, D. (2015). A License Scheme for a Global Federated Language Service Infrastructure. En *WLSI 2015: The Second International Workshop on Worldwide Language Service Infrastructure*, Kyoto, 22-23 de enero.

Comisión Europea. (2007). *Final report from high level group on multilingualism*.

Comisión Europea. (2008). *Multilingualism: an asset for Europe and a shared commitment. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions* COM (2013) 499.

Comisión Europea, Directorate-General for the Internal Market and Services, Meeûs d'Argenteuil, J., Triaille, J. y Francquen, A. (2014). *Study on the legal framework of text and data mining (TDM)*, Publications Office. https://data.europa.eu/doi/10.2780/1475

Consejo de Europa. (2008a). *Council conclusions of 22 May 2002 on multilingualism* (2008/c140/10).

Consejo de Europa. (2008b). *Council resolution of 21 November 2008 on a European strategy for multilingualism*. OJ C 320 16.12.2008. 1–3.

Consejo de Europa. (2010). *Minority language protection in Europe : into a new decade*.

Consejo de Europa. (2014). *Conclusions on multilingualism and the development of language competences*.

Consejo de Europa. (2019). *EUROPEAN CHARTER FOR REGIONAL OR MINORITY LANGUAGES. New technologies, new social media and the European Charter for Regional or Minority Languages. Report for the Committee of Experts*.

Consejo de Europa. (2020). *Languages Covered by the European Charter for Regional or Minority Languages*.

Currey A., y Heafield, K. (2019). Zero-Resource Neural Machine Translation with Monolingual Pivot Data. En *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 99–107.

Danet, B. (2020). *Cyberpl@y: Comunicating online*. Routledge.

Do Campo Bayón, M. y Sánchez-Gijón, P. (Pendiente de publicación). Evaluating NMT using the non-inferiority principle. En *Natural Language Engineering*. Cambridge.

Do Campo Bayón, M. y Sánchez-Gijón, P. (2019). *La traducción automática dentro del contexto de una lengua minorizada. ¿Qué tipo de motor se adapta mejor al caso especial del gallego?* <https://ddd.uab.cat/record/210862> [Consulta: 6 junio de 2023].

Do Carmo, F., Orasan, C., Caro Quintana, R., Saadany, H. y Zilio, L. (2023). Analysing Mistranslation of Emotions in Multilingual Tweets by Online MT Tools. En *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 275-284. Junio de 2023, Tampere, Finlandia.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. En *Proceedings of the 2nd Human Language Technologies Conference (HLT-02)*, 128-132. San Diego, CA, EE.UU.

Doherty, S. (2017). Issues in human and automatic translation quality assessment. En Kenny, D. (Ed.), *Human issues in translation technology*, 131-148. Routledge.

Dreyer, M. y Marcu, D. (2012). HyTER: Meaning-equivalent semantics for translation evaluation. En *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologi*es, 162–171. Montreal, Canadá

EUROPEAN LANGUAGE INDUSTRY SURVEY. (2023). *Trends, expectations and concerns of the European language industry*. ELIS Research.

Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia. (2011). *La protección de las lenguas minoritarias en Europa: hacia una nueva década.*

Explotion AI. (2016). *Spacy: industrial-strenght NLP*. Disponible en: https://spacy.io.

Fadaee, M. y Monz, C. (2018). Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 436–446.

Fadaee, M.; Bisazza, A. y Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *arXiv*.

Fernandez de Arroyabe Olaortua, A., Eguskiza Sesumaga, L. y Lazkano Arrillaga, I. (2020). The future of minority languages on the Internet and young prosumers. The Basque case. En *Cuadernos. Info*, *46*, 367–376. https://doi.org/10.7764/cdi.46.1411

Fernández, M. y Rodríguez, X.. (2016). Avaliación de recursos para a normalización lingüística: as tecnoloxías e os recursos lingüísticos. En *XVIII Encontros para a normalización lingüística*. Consello da Cultura Galega. [vídeo] https://www.youtube.com/watch?v=668GoMhsjPg. Fecha de consulta: 25 de agosto de 2023.

Ferré-Pavia, C., Zabaleta, I., Gutierrez, A., Fernandez-Astobiza, I y Xamardo, N. (2018). Internet and Social Media in European Minority Languages: Analysis of the Digitalization Process. En *International Journal of Communication* (Vol. 12). http://ijoc.org.

Forcada, M. (2017). Making sense of neural machine translation. En *Translation spaces* 6(2), 291-309.

Gamallo, P., Bardanca, D., Pichel, J. R., García, M., Rodríguez-Rey, S. y de-Dios-Flores, I. (2023). NOS-MT-OpenNMT-es-gl. Disponible: https://huggingface.co/proxectonos/NOS-MT-OpenNMT-es-gl. Fecha de consulta: 14 de septiembre de 2023.

García Mateo, C. y Arza Rodríguez, M. (2012). *O idioma galego na era dixital*. Berlín: Springer Heidelberg.

González, D. y Rico, C. (2021). POSEDITrad: la traducción automática y la posedición para la formación de traductores e intérpretes. En *Revista Digital de Investigación en Docencia Universitaria*, 15, 1-14.

González Martínez, M. (2002). El paper de la traducció en la normalització de la llengua gallega. En Díaz-Fouces, O., Gónzalez, M., Costa Carreras, J. (ed). *Traducció i dinámica sociollingüística*. Llibres de l'index.

González Martínez, M. y Veiga Días, M. (2007). A recepción da traducción de Harry Potter ás linguas romances da Península. En González Fernández, H. y Lama López, M. X. (Coords.). *Mulleres en Galicia. Galicia e os otros pobos da península: actas VII Congreso Internacional de Estudos Galegos* (2), 971-084.

González-Bailón, S., Borge-Holthoefer, J., Rivero, A. y Moreno, Y. (2012). The Dynamics of Protest Recruitment through an Online Network. En *Scientific Reports*, 2(1), 1-8.

Görög A. (2014). Quantifying and benchmarking quality: The TAUS Dynamic Quality Framework. En *Revista Tradumàtica: tecnologies de la traducció, Traducció i qualitat 12*, 443-454. ISSN: 1578–7559. Disponible en: http://revistes.uab.cat/tradumatica.

Gouadec, D. (2010). Quality in Translation. En *Handbook of Translation Studies*, 1, 270-275. John Benjamins Publishing Company.

Goyal, V., Kumar, S. y Sharma, D.M. (2020). Efficient neural machine translation for low-resource languages via exploiting related languages. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 162-168. Online, 5–10 de julio de 2020.

Graham Y., Baldwin T., Moffat A. y Zobel J. (2015). Can machine translation systems be evaluated by the crowd alone. En *Natural Language Engineering* 23(1), 3–30.

Grbić, N. (2008). Constructing interpreting quality. En *Interpreting*, 10(2), 232-257.

Gu, J., Wang, Y., Chen, Y., Cho, K. y Li, V.O.K. (2018). Meta-learning for low-resource neural machine translation. In *arXiv*

Guerberof-Arenas, A. y Toral, A. (2022). Creativity in translation: machine translation as a constraint for literary texts. En *Translation Spaces*. John Benjamins Publishing Company.

Guibault, L. y Margoni, T. (2015). Legal Aspects of Open Access to Publicly Funded Research. En *OECD, Enquiries into Intellectual Property's Economic Impact*, 373-414.

Guibault, L. y Wiebe, A. (Eds.). (2013). *Safe to be open. Study on the protection of research data and recommendations for access and usage*. Göttingen: Universitätsverlag Göttingen.

Guzmán, R. (2007). Advanced automatica MT postedition using regular expressions. En *Multilingual* 18(6): 49-52.

Han, A. L. F., Wong D. F. y Chao L. S. (2012). LEPOR: A robust evaluation metric for machine translation with augmented factors. En *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*: Posters, 441–450. Mumbai, India.

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. En *arXiv preprint arXiv:1803.05567*.

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M. y Wright, J. (2014). The Language Application Grid En *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*

IGE. (2018). *Enquisa estructural a fogares*.

Imankulova, A., Sato, T. y Komachi, M. (2017). Improving low-resource neural machine translation with filtered pseudo-parallel corpus. En *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, 70-78. Taipei, Taiwan, del 27 de noviembre al 1 de diciembre de 2017.

Instrumento de ratificación de la Carta Europea de las Lenguas Regionales o Minoritarias, hecha en Estrasburgo el 5 de noviembre de 1992. Boletín Oficial del Estado, núm. 222, de 15 de septiembre de 2001, 34733 - 34749.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T. y Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. En *Proceedings of ACL 2018, System Demonstrations*, 116-121. Julio de 2028, Melbourne, Australia.

Kardos, G. (2020). The European Charter for Regional or Minority Languages. En *Hungarian Yearbook of International Law and European Law*, 7(1), 259–269. https://doi.org/10.5553/hyiel/266627012019007001015.

Keller, P., Margoni, T., Rybicka, K., y Tarkowski, A. (2014). Re-Use of Public Sector Information in Cultural Heritage Institutions. En *IFOSS Law Review*, 6 (1), 1-9.

Koby, G. S., Fields, P., Hague, D., Lommel, A. y Melby, A. (2014). Defining Translation Quality. En *Revista Tradumàtica: tecnologies de la traducció*, 12, 413-420.

Koehn, P. (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*. En Proceedings of the MT Summit X, 79-86.

Kudo, T. y Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. En *arXiv preprint arXiv:1808.06226*.

Lacruz, I., Denkowski, M. y Lavie, A. (2014). Cognitive demand and cognitive effort in post-editing. En *Proceedings of the Third Workshop on Post-Editing Technology and Practice*. 11th Conference of the Association for Machine Translation in the Americas. Vancouver, BC, Canadá.

Lakew, S. M., Federico, M., Negri, M., Turchi, M. (2018). Multilingual neural machine translation for low-resource languages. En *Italian Journal of Computational Linguistics*, 4, 11–25.

Lankford, S., Afli, H. y Way, A. (2022). Human Evaluation of English–Irish Transformer-Based NMT. En *Information 2022*, 13 (7), 309. https://doi.org/10.3390/ info13070309

Leppänen, S. y Kytölä, S. (2016). Twitter, language and identity: the *hashtag* as an indexing and discursive practice. En *Digital Discourse*, 81-102. Palgrave Macmillan.

Leusch, G., Ueffing, N. y Ney, H. (2003). A novel string-to-string distance measure with applications to machine translation evaluation. En *Proceedings of MT Summit IX*, Nueva Orleans, LA, EE.UU.

Leusch, G., Ueffing, N. y Ney, H. (2006). CDER: Efficient MT evaluation using block movements. En *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italia.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. En *Soviet Physics – Doklady* 10(8), 707–710.

Liu, Q., Kusner, M. y P. Blunsom. (2021). Counterfactual Data Augmentation for Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 187–197.

Lohar, P., Popović, M., Alfi, H. y Way, A. (2019). A systematic comparison between SMT and NMT on translating user-generated content. En 2*0th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*, 7 - 13 de abril de 2019, La Rochelle, Francia.

Lommel, A. (2013). Best practices for translation and localization of social media. En *Galaxy Newsletter*.

Lommel, A., Uszkoreit, H. y Burchardt, A. (2014). MQM: a framework for declaring and describing translation quality metrics. En *Revista Tradumàtica: tecnologies de la traducció*, 12, 455-463.

Lommel, A. y Melby, A. (2018). Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). En *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*. Boston.

Luong, M. y Manning, C. (2015). Stanford Neural Machine Translation Systems for Spoken Language Domains. En *Proceedings of the International Workshop on Spoken Language Translation 2015*. Da Nang, Vietnam.

Martínez-Cortés, J. P. y Paricio-Martín, S. J. (2010). New ways to revitalise minority languages: the impact of the internet in the case of Aragonese, En *Digithum*, 12 (mayo).

Massardo I., Van der Meer J., O'Brien S., Hollowood F., Aranberri N. and Drescher K. (2016). *MT Post-Editing Guidelines*. The Netherlands: TAUS Signature Editions.

Melamed, I., Green, R. y Turian, J. (2003). Precision and recall of machine translation. En *Proceedings of the HLT-NAACL 2003*. Edmonton, Canadá.

Miceli Barone, A. V., Haddow, B., Germann, U. y Sennrich, R. (2017). Regularization techniques for fine-tuning in neural machine translation. En *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1489–1494. Copenhage, Dinamarca, 7–11 de septiembre de 2017.

Miceli Barone, A. V., Helcl, J., Sennrich, R., Haddow, B. y Birch, A. (2017). Deep architectures for Neural Machine Translation. En *Proceedings of the Second Conference on Machine Translation:* 99-107.

Molina Nadal, A. (2020). Ensayos clínicos de no inferioridad. En *FMC - Formación Médica Continuada en Atención Primaria*, 27(7), 345-348.

Monteagudo, H. (2019). Política lingüística en Galicia: de la normalización sin conflicto al conflicto desnormalizador. En Guiralt Latorre, J. y Nagore Laín, F. *La NORMALIZACIÓN social de las lenguas minoritarias: experiencias y procedimientos para la salvaguarda de un patrimonio inmaterial*. Zaragoza: Prensas de la Universidad de Zaragoza.

Naciones Unidas. (1992). *Declaración sobre los Derechos de las Personas pertenecientes a Minorías Nacionales o Étnicas, Religiosas y Lingüísticas*. Resolución 47/135, 18 de diciembre de 1992, disponible en esta dirección: https://www.refworld.org.es/docid/5d7fc5a32.html [Consultado el 2 Septiembre 2023]

Navarro, G. (2001). A guided tour to approximate string matching. En *ACM Computing Surveys* 33(1), 31–88.

Niessen. S., Och. F., Leusch. G. y Ney. H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. En *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Atenas, Grecia.

O'Brien, S. (2022). How to deal with errors in machine translation: Post-editing. En Kenny, D. (Ed). (2022). *Machine translation for everyone: Empowering users in the age of artificial intelligence*. (Translation and Multilingual Natural Language Processing 18). Berlin: Language Science Press. DOI: 10.5281/zenodo.6653406.

Oliver, A. (2020). MTUOC: easy and free integration of NMT systems in professional translation environments. En *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 467–468. Lisboa, Portugal, noviembre de 2020.

Oliver, A., Vázquez, M., Coll-Florit, M., Álvarez, S., Suárez, V., Aventín-Boya, C., Valdés, C., Font, M. y Pardos, A. (2023). TAN-IBE: Neural Machine Translation for the romance languages of the Iberian Peninsula. En *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* 495-497. Tampere, Finlandia, 12 – 15 junio de 2023.

Organización de Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (1960). *Convención relativa a la lucha contra las discriminaciones en la esfera de la enseñanza*, 14 Diciembre 1960, disponible en esta dirección: https://www.refworld.org.es/docid/5204c69f4.html [Consultado el 2 Septiembre 2023]

Organización de Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2001). *Declaración Universal de la UNESCO sobre la Diversidad Cultural*. Adoptada por la 31ª reunión de la Conferencia General de la UNESCO París, 2 de noviembre de 2001.

Organización de Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2003) *Recomendaciones sobre la Promoción y el Uso del Plurilingüismo y el Acceso Universal al Ciberespacio*. Disponible en: https://en.unesco.org/sites/default/files/spa_-_recommendation_concerning_the_promotion_and_use_of_multilingualism_and_universal_access_to_cyberspace.pdf (Consultado el 2 de septiembre de 2023).

Ortega, J. E., de-Dios-Flores, I., Campos, J. R. P., y Gamallo, P. (2022). A neural machine translation system for Spanish to Galician through Portuguese transliteration. En *Demo presented at the 15th International Conference on Computational Processing of Portuguese (PROPOR 2022)*.

Papacharissi, Z. (2010). *A networked self: Identity, community, and culture on social network sites*. Routledge.

Papineni, K., Roukos, S., Ward, T., y Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting on association for computational linguistic*s (ACL), 311-318.

Park, C., Yang, Y., Park, K., y Lim, H. (2020). Decoding Strategies for Improving Low-Resource Machine Translation. Electronics 2020, 9(10), 1562. https://doi.org/10.3390/ELECTRONICS9101562

Parlamento Europeo. (2020). *European Day of Languages: Digital survival of lesser-used languages*. EPRS.

Parlamento Europeo. (2020). *Linguistic diversity on the internet. Assessment of the contribution of machine translation: final study.*

Pasikowska-Schnass, M. (2016). *Regional and minority languages in the European Union*. EPRS.

Pérez-Ortiz, J. A., Forcada, M. L. y Sánchez Martínez, F. (2022). How neural machine translation Works. En Kenny, Dorothy (ed). *Machine translation for everyone: Empowering users in the age of artificial intelligence*. (Translation and Multilingual Natural Language Processing 18). Berlin: Language Science Press. DOI: 10.5281/zenodo.6653406.

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 mayo, European Language Resources Association (ELRA).

Popovi´c, M. (2015). CHRF: Character n-gram F-score for automatic MT evaluation. En *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392-395. Lisboa, Portugal.

Popovi´c M. (2018). Error classification and analysis for machine translation quality assessment. En Moorkens, J., Castilho, S., Gaspari, F. y Doherty, S. (Eds). *Translation Quality Assessment*. *From Principles to Practice*. Cham, Suiza: Springer.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. En *Proceedings of the Third Conference on Machine Translation: Research Papers*, 1, 186-191. https://doi.org/10.18653/v1/W18-64019

Ranathunga, S., Lee, E. Annie, S., Prifti, M. y Shekhar, R. (2023). Neural Machine Translation for Low-Resource Languages: A Survey. En *arXiv e-prints*.

Rapp, R. (2021) Similar Language Translation for Catalan, Portuguese and Spanish Using Marian NMT. En *Proceedings of the Sixth Conference on Machine Translation*, 292–298, Online. Association for Computational Linguistics.

Rehm, G. (Ed.). (2022). *European Language Grid: A Language Technology Platform for Multilingual Europe. Cognitive Technologies*. Springer, Cham, Switzerland, January 2023

Rei, R., Stewart, C., Farinha, A. C., y Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. En *Proceeding of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2685–2702. Association for Computational Linguistics

Rico, C. (2017). La formación de traductores en traducción automática. En *Revista Tradumàtica*, 15, 75-96.

Rodriguez-Doncel, V. y Labropoulou, P. (2015). RDF Representation of Licenses for Language Resources. En *4th Workshop on Linked Data in Linguistics: Resources and Applications, ACL-IJCNLP 2015*, Beijing, China.

Rossi, C. y Carré, A. (2022). How to choose a suitable neural machine translation solution. En Kenny, D. (Ed). *Machine translation for everyone: Empowering users in the age of artificial intelligence*. (Translation and Multilingual Natural Language Processing 18). Berlin: Language Science Press. DOI: 10.5281/zenodo.6653406.

Sánchez-Gijón, P. (2016). La posedición: hacia una definición competencial del perfil y una descripción multidi-mensional del fenómeno. En *Sendebar*, 27, 151-162.

Sánchez-Gijón, P., Moorkens, J., y Way, A. (2019). Post-editing neural machine translation versus translation memory segments. En *Machine Translation*, 33(1-2), 31-59.

Sánchez-Gijón, P. y Palenzuela-Badiola, L. (2023). Analysis And Evaluation of ChatGPT-Induced HCI Shifts In The Digitalised Translation Process. En *Proceedings of the International Conference HiT-IT 2023*, 227–267. 7-9 de julio de 2023. Nápoles, Italia.

Sánchez Ramos, M. D. M. y Rico Pérez, C. (2020). *Traducción automática: conceptos clave, procesos de evaluación y técnicas de posedición*. Granada, Editorial Comares.

Senn, S. (2000). Consensus and Controversy in Pharmaceutical Statistics. En J*ournal of the Royal Statistical Society. Series D (The Statistician)*, 49(2), 135–176.

Sennrich, R., Haddow, B., & Birch, A. (2016a). Neural Machine Translation of Rare Words with Subword Units. En *arXiv preprint arXiv:1508.07909*.

Sennrich, R., Haddow, B. y Birch, A. (2016b). Improving Neural Machine Translation Models with Monolingual Data. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 86– 96. Berlin, Alemania.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. y Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. En *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Boston Marriott, Cambridge, Massachusetts, EE.UU.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M. y Cristianini, N. (2009). Estimating the Sentence Level Quality of Machine Translation Systems. En *EAMT09*, 28-37. Barcelona, España.

Specia L., Raj D. & Turchi M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation* 24, 39–50. Springer Science+Business Media B.V. doi:10.1007/s10590-010-9077-2.

Specia, L., Shah, K., De Souza, J. G. C. y Cohn, T. (2013). QuEst – A translation quality estimation framework. En *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 79-84. Sofia, Bulgaria.

Stymne, S. (2011). Pre-and Postprocessing for Statistical Machine Translation into Germanic Languages. En *Proceedings of the ACL 2011 Student Session*, 12-17.

Tomás, J., Mas, J. A. y Casacuberta, F. (2003). A quantitative method for machine translation evaluation. En *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable?*, Budapest, Hungría.

Toral, A., Castilho, S., Hu, K. y Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. En *Proceedings of the Third Conference on Machine Translation (WMT)*, Volumen 1: 113-123. Research Papers, Association for Computational Linguistics, Brussels, Belgium.

Torres-Hostench, O., Cid-Leal, P., Presas, M. (Coords.). (2016). *El uso de traducción automática y posedición en las empresas de servicios lingüísticos españolas: informe de investigación ProjecTA 2015*. Bellaterra.

Torres-Hostench, Olga. (2023). Europe, multilingualism and machine Translation. En Kenny, D. (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 187–207. Berlin: Language Science Press.

Translation Automation User Society. (2012). Machine Translation Post-editing guidelines. https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines

Tunes da Silva, G., Logan, B. y Klein, J. P. (2009). Methods for equivalence and noninferiority testing. En *Biol Blood Marrow Transplant*, 15(1 Suppl), 120-127.

UNESCO. 2003. *RECOMENDACIÓN SOBRE LA PROMOCIÓN Y EL USO DEL PLURILINGÜISMO Y EL ACCESO UNIVERSAL AL CIBERESPACIO UNESCO.* París.

Van der Werff, T., van Noord, R. y Toral, A. (2022). Automatic Discrimination of Human and Neural Machine Translation: A Study with Multiple Pre-Trained Models and Longer Context. En *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 161-170, Ghent, Bélgica. European Association for Machine Translation.

Van Edgom, G. y Pluymaekers, M. (2019). Why go the extra mile? How different degrees of post-editing affect perceptions of texts, senders and products among end users. En *The Journal of Specialised Translation, 31, 158-176.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention is all you need. En *Advances in Neural Information Processing Systems*, 5998–6008.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. y Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. En *ArXiv*:1609.08144.

Zappavigna, M. (2012). *Discourse of Twitter and social media: How we use language to create affiliation on the web*. Bloomsbury Publishing.

Zhang, B., Nagesh, A., y Knight, K. (2020). Parallel Corpus Filtering via Pre-trained Language Models. En ArXiv, abs/2005.06166.

Zhou, L., Lin, C.-Y. y Hovy, E. (2006). Re-evaluating machine translation results with paraphrase support. En *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sydney, Australia.

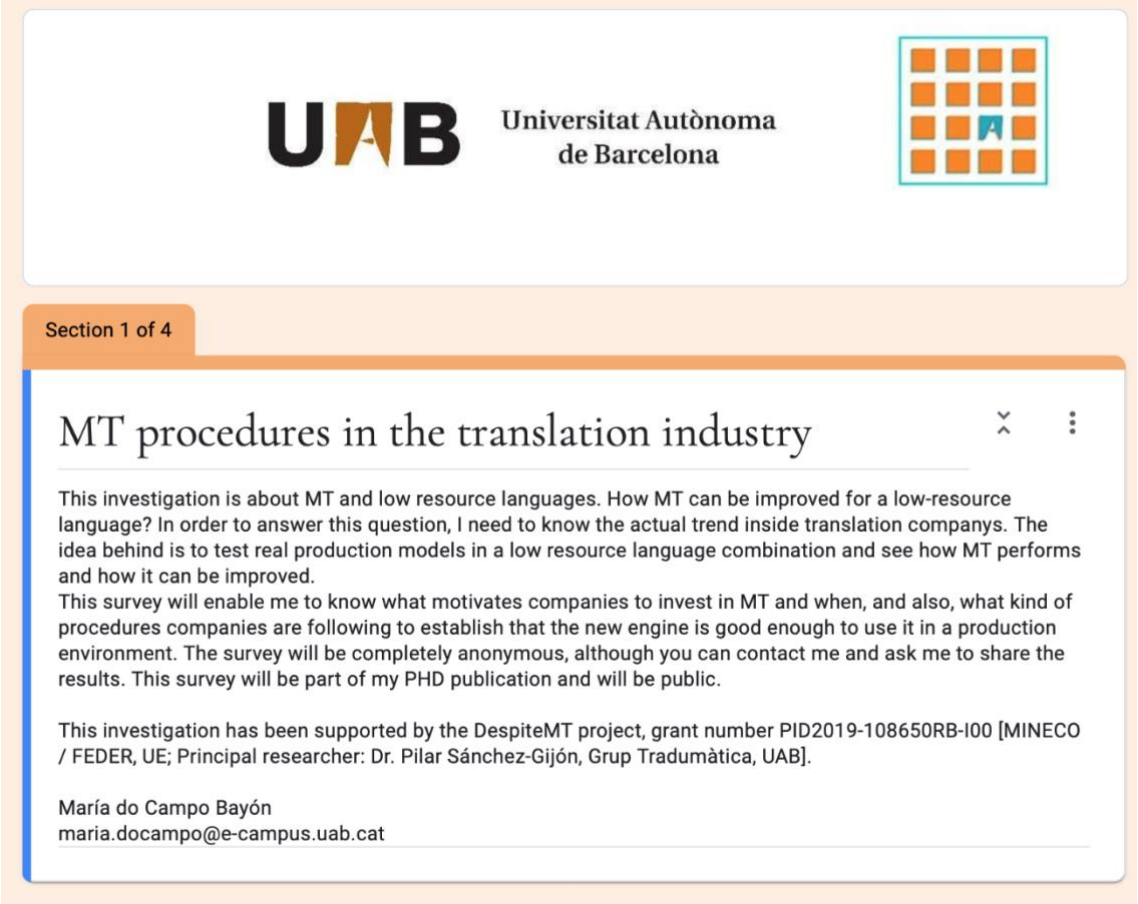
Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D. y Zhao, T. (2008). Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. En *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008),* 1121-1128. Manchester, Reino Unido.

Zimbra, D., Abbasi, A., Zeng, D. y Chen, H. (2018). The state-of-the-art in twitter sen- timent analysis: A review and benchmark evaluation. En *ACM Trans. Manage. Inf. Syst.* 9, 5:1–5:29.

Zoph, B., Yuret, D., May, J. y Knight, K. (2016) *Transfer learning for low-resource neural machine translation*. En *arXiv:1604.02201*.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., y Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R* (Vol. 574). Springer, New York.

Anexo I: encuesta sobre procesos de TA dentro de la industria de la traducción



Universitat Autònoma de Barcelona

Section 1 of 4

MT procedures in the translation industry

This investigation is about MT and low resource languages. How MT can be improved for a low-resource language? In order to answer this question, I need to know the actual trend inside translation companys. The idea behind is to test real production models in a low resource language combination and see how MT performs and how it can be improved.
This survey will enable me to know what motivates companies to invest in MT and when, and also, what kind of procedures companies are following to establish that the new engine is good enough to use it in a production environment. The survey will be completely anonymous, although you can contact me and ask me to share the results. This survey will be part of my PHD publication and will be public.

This investigation has been supported by the DespiteMT project, grant number PID2019-108650RB-I00 [MINECO / FEDER, UE; Principal researcher: Dr. Pilar Sánchez-Gijón, Grup Tradumàtica, UAB].

María do Campo Bayón
maria.docampo@e-campus.uab.cat

General information about your company

Description (optional)

How many employees do you have? *

Please consider all employees of your company: in-house translators, project managers, vendors, etc.

- ○ Less than 25
- ○ From 26 to 50
- ○ From 51 to 75
- ○ From 76 to 100
- ○ More than 100

Where are your offices based? *

You can select more than one option.

- ☐ Europe
- ☐ America
- ☐ Asia
- ☐ Africa
- ☐ Other...

Do you have a specific department for MT? *

- ○ Yes
- ○ No
- ○ Other...

MT procedures

Description (optional)

Training a new engine is usually motivated by... *

You can select more than one option.

- [ ] Client
- [ ] Translation Fee
- [ ] Time
- [ ] Combination of all
- [ ] Other...

Which type of resources do you use to train the engine? *

- [ ] Client's TM
- [ ] Corpus from specific genre
- [ ] General corpus
- [ ] Other...

Have you established the minimal amount of data you will need to create a new engine? *

- ○ Yes, for one or more language pairs
- ○ No

If you have established the minimal amount of data, please indicate it:

Short answer text

Which type of process do you use to test the engine? *

Long answer text

Which kind of indicators do you use? *

- ☐ Automatic Evaluation
- ☐ Human Evaluation
- ☐ Edit Distance
- ☐ Postedit Effort
- ☐ Mix of them
- ☐ Other...

Which kind of comparisons do you do? *

- [ ] Specific vs. specific MT engine (different engine provider)
- [ ] Generic vs. generic MT engine (different engine provider)
- [ ] Specific vs. generic MT (same type of engine)
- [ ] Neuronal vs. PB MT engine
- [ ] Neuronal vs. PB vs. RB MT engine
- [ ] Human translation
- [ ] None
- [ ] Other...

What makes you adopt the AT engine or reject it? *

Long answer text

What's your minimal threshold to consider an engine available to use in a production environment? What is the process to stablish quality?

As the answer may vary depending on the type of project, consider this two possible scenarios

What's your minimal threshold to translate low impact/visibility technical documentation?

Short answer text

What's your minimal threshold to translate a high impact/visibility project?

Long answer text

Why would you perform improvements to an MT engine that is already working? *

- [ ] Posteditors performance
- [ ] Client's needs
- [ ] Fee
- [ ] I wouldn't perform improvements to an MT engine that is already working
- [ ] Other...

Is translator feedback about the raw MT output taken into account? *

- ( ) Yes
- ( ) No
- ( ) Other...

Is COVID-19 affecting any of your decisions towards investing in MT? *

- ( ) Yes
- ( ) No

How is it affecting your decisions?

Short answer text

Anexo II: *script* de preprocesamiento del corpus

```
MTUOC: /MTUOC


preprocess_type: sentencepiece

#one of smt sentencepiece subwordnmt


corpus:

from_train_val: False

train_corpus:

val_corpus:

valsize: 5000

evalsize: 5000

SLcode3: spa

SLcode2: es

TLcode3: gal

TLcode2: gl


SL_DICT: /MTUOC/spa.dict

TL_DICT: /MTUOC/gal.dict

#state None or null.dict if not word form dictionary available for
that languages


SL_TOKENIZER: MTUOC_tokenizer_spa.py

TOKENIZE_SL: False

TL_TOKENIZER: MTUOC_tokenizer_gal.py
```

```
TOKENIZE_TL: False


###PREPARATION

REPLACE_EMAILS: True

EMAIL_CODE: "@EMAIL@"

REPLACE_URLS: True

URL_CODE: "@URL@"


TRAIN_SL_TRUECASER: True

TRUECASE_SL: True

SL_TC_MODEL: auto

#if auto the name will be tc.+SLcode2


TRAIN_TL_TRUECASER: True

TRUECASE_TL: True

TL_TC_MODEL: auto

#if auto the name will be tc.+TLcode2


CLEAN: True

MIN_TOK: 1

MAX_TOK: 80


MIN_CHAR: 1

MAX_CHAR: 1000
```

```
#SENTENCE PIECE and SUBWORD NMT

bos: <s>

#<s> or None

eos: </s>

#</s> or None

JOIN_LANGUAGES: True

SPLIT_DIGITS: True

VOCABULARY_THRESHOLD: 50


#SMT

REPLACE_NUM: True

NUM_CODE: "@NUM@"



#SENTENCE PIECE

CONTROL_SYMBOLS: ""

USER_DEFINED_SYMBOLS:
"<tag0>,<tag1>,<tag2>,<tag3>,<tag4>,<tag5>,<tag6>,<tag7>,<tag8>,<ta
g9>,<tag10>,</tag0>,</tag1>,</tag2>,</tag3>,</tag4>,</tag5>,</tag6>
,</tag7>,</tag8>,</tag9>,</tag10>,<tag0/>,<tag1/>,<tag2/>,<tag3/>,<
tag4/>,<tag5/>,<tag6/>,<tag7/>,<tag8/>,<tag9/>,<tag10/>,"

SP_MODEL_PREFIX: spmodel

MODEL_TYPE: bpe

#one of unigram, bpe, char, word

VOCAB_SIZE: 64000

CHARACTER_COVERAGE: 1.0
```

```
CHARACTER_COVERAGE_SL: 1.0

CHARACTER_COVERAGE_TL: 1.0

INPUT_SENTENCE_SIZE: 5000000


#SUBWORD NMT

LEARN_BPE: True

JOINER: "@@"

#use one of ▪ or @@

NUM_OPERATIONS: 85000

APPLY_BPE: True

BPE_DROPOUT: True

BPE_DROPOUT_P: 0.1



#GUIDED ALIGNMENT

#TRAIN CORPUS

GUIDED_ALIGNMENT: True

ALIGNER: eflomal

#one of eflomal, fast_align, simalign, awesome

DELETE_EXISTING: True

SPLIT_LIMIT: 1000000

#For efomal, max number of segments to align at a time


#VALID CORPUS

GUIDED_ALIGNMENT_VALID: True
```

```
ALIGNER_VALID: eflomal

#one of eflomal, fast_align, simalign, awesome

DELETE_EXISTING_VALID: True


#simalign specific options

simalign:

  device: cuda

  matching_method: i

  #"a": "inter"; "m": "mwmf"; "i": "itermax"; "f": "fwd"; "r": "rev";
mai


#awesome specific options

awesome:

    finetune: False

    finetune_limit: 1000000

    finetuned_dir: ./finetuned_model

    finetune_initial_model: bert-base-multilingual-cased

    finetune_device: cuda

    finetune_cuda_visible_devices: 0


    align_model: bert-base-multilingual-cased

    align_device: cuda

    align_cuda_visible_devices: 0


VERBOSE: True
```

```
LOG_FILE: process.log

DELETE_TEMP: True
```

Anexo III: distribución de los tuits en los cuestionarios de la prueba piloto

Cuestionario 1

Número de pregunta	Tuit	Común con otro cuestionario
1	TAPM4	Sí, cuestionario 2
2	TAPM6	Sí, cuestionario 2
3	TAPM5	Sí, cuestionario 2
4	TAPL2	Sí, cuestionario 2
5	TOFC1	Sí, cuestionario 2
6	TOPM1	
7	TAPC6	
8	TOPC3	Sí, cuestionario 4
9	TOFL6	Sí, cuestionario 4
10	TOFC2	Sí, cuestionario 4
11	TOPC6	Sí, cuestionario 4
12	TAPC4	
13	TAPL1	
14	TOFC4	
15	TAFL4	
16	TOPL4	Sí, cuestionario 4
17	TAFC6	
18	TAPM2	
19	TOFC6	
20	TAPL6	

Distribución: 11 TA / 9 TO

Leyenda

TO = Texto original redactado en gallego

TA = Traducción automática al gallego

FC = Frase corta

FL = Frase Larga

PC = Párrafo frases cortas

PL = Párrafo con frases largas

PM = Párrafo con mezcla de frases largas y cortas

Cuestionario 2

Número de pregunta	Tuit	Común con otro cuestionario
1	TAPM4	Sí, cuestionario 1
2	TAPM6	Sí, cuestionario 1
3	TAPM5	Sí, cuestionario 1
4	TAPL2	Sí, cuestionario 1
5	TOFC1	Sí, cuestionario 1
6	TAFL6	
7	TOPM4	Sí, cuestionario 3
8	TAPL3	
9	TAPL5	
10	TOPL2	Sí, cuestionario 3
11	TAFL5	
12	TAPC2	
13	TOPL5	Sí, cuestionario 3
14	TAPM3	
15	TAFC4	
16	TOFL1	Sí, cuestionario 3
17	TOPM5	Sí, cuestionario 3
18	TAPC3	
19	TOFL5	
20	TOPC2	

Distribución: 12 TA / 8 TO

Leyenda

TO = Texto original redactado en gallego

TA = Traducción automática al gallego

FC = Frase corta

FL = Frase Larga

PC = Párrafo frases cortas

PL = Párrafo con frases largas

PM = Párrafo con mezcla de frases largas y cortas

Cuestionario 3

Número de pregunta	Tuit	Común con otro cuestionario
1	TOPM4	Sí, cuestionario 2
2	TOPL2	Sí, cuestionario 2
3	TOPL5	Sí, cuestionario 2
4	TOFL1	Sí, cuestionario 2
5	TOPM5	Sí, cuestionario 2
6	TOPL3	
7	TOFL2	
8	TOPL6	
9	TAFC1	
10	TAFL3	
11	TAPM1	
12	TAFC5	
13	TOPM3	Sí, cuestionario 4
14	TOPM6	Sí, cuestionario 4
15	TOFL4	Sí, cuestionario 4
16	TAPL4	Sí, cuestionario 4
17	TAFC2	Sí, cuestionario 4
18	TAPC1	
19	TOPC1	
20	TOFC5	

Distribución: 7 TA / 13 TO

Leyenda

TO = Texto original redactado en gallego

TA = Traducción automática al gallego

FC = Frase corta

FL = Frase Larga

PC = Párrafo frases cortas

PL = Párrafo con frases largas

PM = Párrafo con mezcla de frases largas y cortas

Cuestionario 4

Número de pregunta	Tuit	Común con otro cuestionario
1	TOPM3	Sí, cuestionario 3
2	TOPM6	Sí, cuestionario 3
3	TOFL4	Sí, cuestionario 3
4	TAPL4	Sí, cuestionario 3
5	TAFC2	Sí, cuestionario 3
6	TOPC3	Sí, cuestionario 1
7	TOFL6	Sí, cuestionario 1
8	TOFC2	Sí, cuestionario 1
9	TOPC6	Sí, cuestionario 1
10	TOPL4	Sí, cuestionario 1
11	TOPC4	
12	TAPC5	
13	TOPC5	
14	TOPL1	
15	TOFC3	
16	TOFL3	
17	TOPM2	
18	TAFL2	
19	TAFC3	
20	TAFL1	

Distribución: 6 TA / 14 TO

Leyenda

TO = Texto original redactado en gallego

TA = Traducción automática al gallego

FC = Frase corta

FL = Frase Larga

PC = Párrafo frases cortas

PL = Párrafo con frases largas

PM = Párrafo con mezcla de frases largas y cortas

Anexo IV: análisis bivariado descriptivo comparando la percepción con las distintas variables

	Non *N=468*	Si *N=1712*	p-value	N
Cuestionario:			<0.001	2180
1	59 (20.3%)	231 (79.7%)		
2	65 (27.1%)	175 (72.9%)		
3	58 (23.2%)	192 (76.8%)		
4	28 (12.2%)	202 (87.8%)		
5	43 (16.5%)	217 (83.5%)		
6	63 (30.0%)	147 (70.0%)		
7	50 (25.0%)	150 (75.0%)		
8	53 (17.7%)	247 (82.3%)		
9	49 (24.5%)	151 (75.5%)		
Edad:			<0.001	2180
De 18 a 25	44 (14.7%)	256 (85.3%)		
De 25 a 35	124 (24.8%)	376 (75.2%)		
De 35 a 45	90 (17.0%)	440 (83.0%)		
De 45 a 55	118 (25.7%)	342 (74.3%)		
De 55 a 65	80 (23.5%)	260 (76.5%)		
Maior de 65	12 (24.0%)	38 (76.0%)		
Competencia lingüística en galego:			0.071	2180
Nivel baixo como falante	19 (27.1%)	51 (72.9%)		
Nivel medio como falante	77 (22.6%)	263 (77.4%)		
Nivel alto como falante	267 (19.8%)	1083 (80.2%)		
Profesional (revisor, profesor de galego, tradutor, etc.)	105 (25.0%)	315 (75.0%)		
Uso o galego:			0.207	2180

	Non	Si	p-value	N
	N=468	*N=1712*		
Non, nunca	22 (31.4%)	48 (68.6%)		
Maioritariamente o castelán	94 (19.2%)	396 (80.8%)		
Na mesma proporción	85 (21.8%)	305 (78.2%)		
Maioritariamente o galego	71 (20.9%)	269 (79.1%)		
Si, sempre	196 (22.0%)	694 (78.0%)		
Uso o galego en RRSS:			0.757	2180
Non, nunca	35 (19.4%)	145 (80.6%)		
Maioritariamente o castelán	102 (20.8%)	388 (79.2%)		
Na mesma proporción	47 (24.7%)	143 (75.3%)		
Maioritariamente o galego	88 (21.0%)	332 (79.0%)		
Si, sempre	196 (21.8%)	704 (78.2%)		
Formación Reglada en galego:			0.098	2180
Non	47 (17.4%)	223 (82.6%)		
Si	421 (22.0%)	1489 (78.0%)		

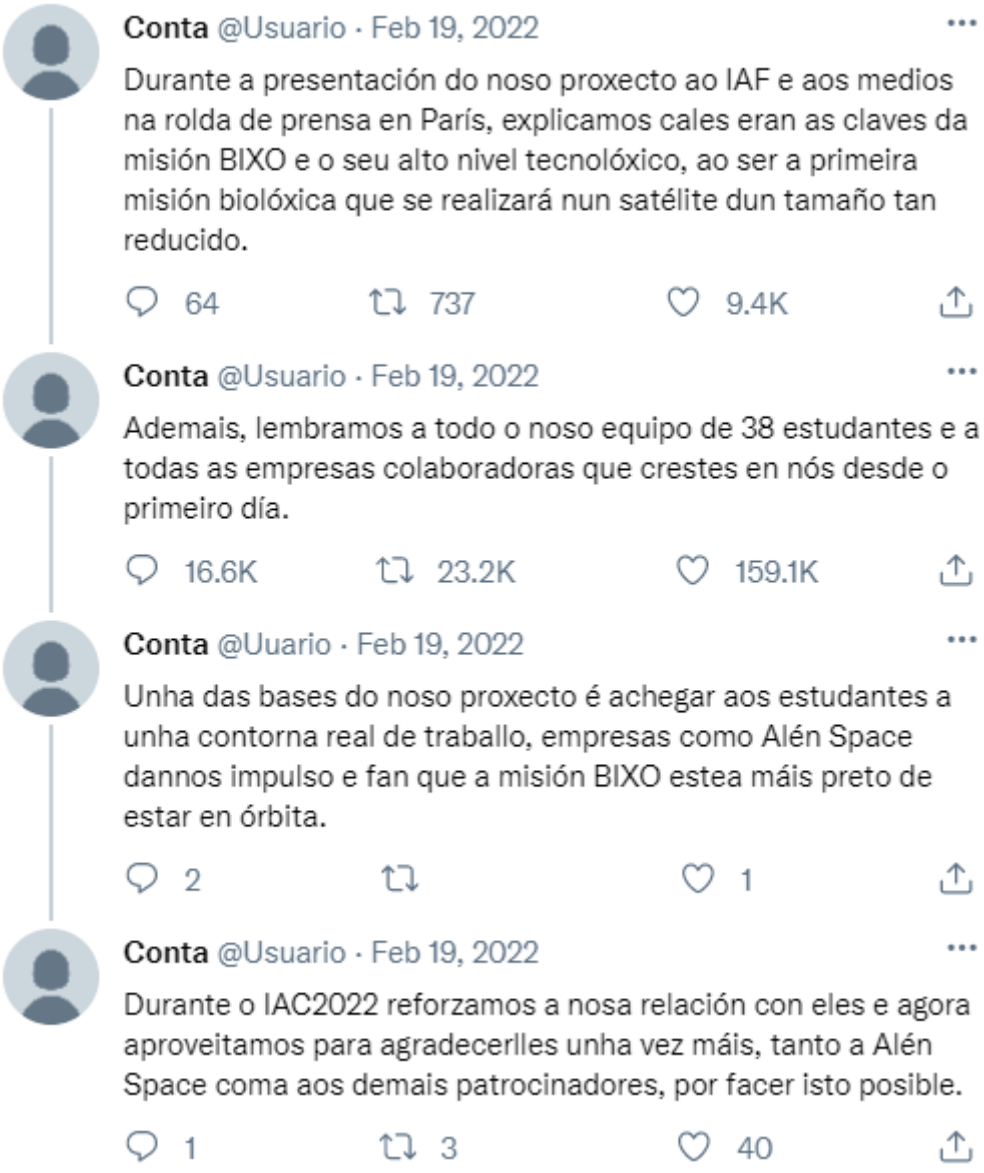


Ilustración 71. AFío1.

Ilustración 72. AFío2.

Ilustración 73. AFío3.

Ilustración 74. AFC1.

Ilustración 75. AFC2.

Ilustración 76. AFC3.

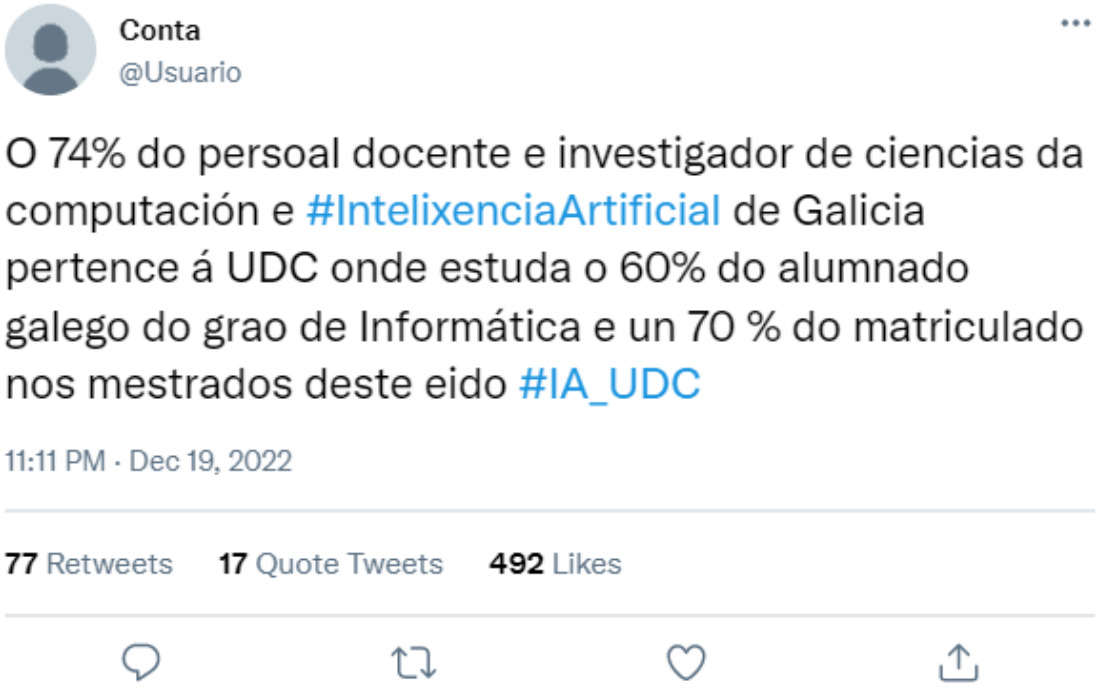


Ilustración 77. AFL1.

Ilustración 78.AFL2.

Ilustración 79. AFL3.

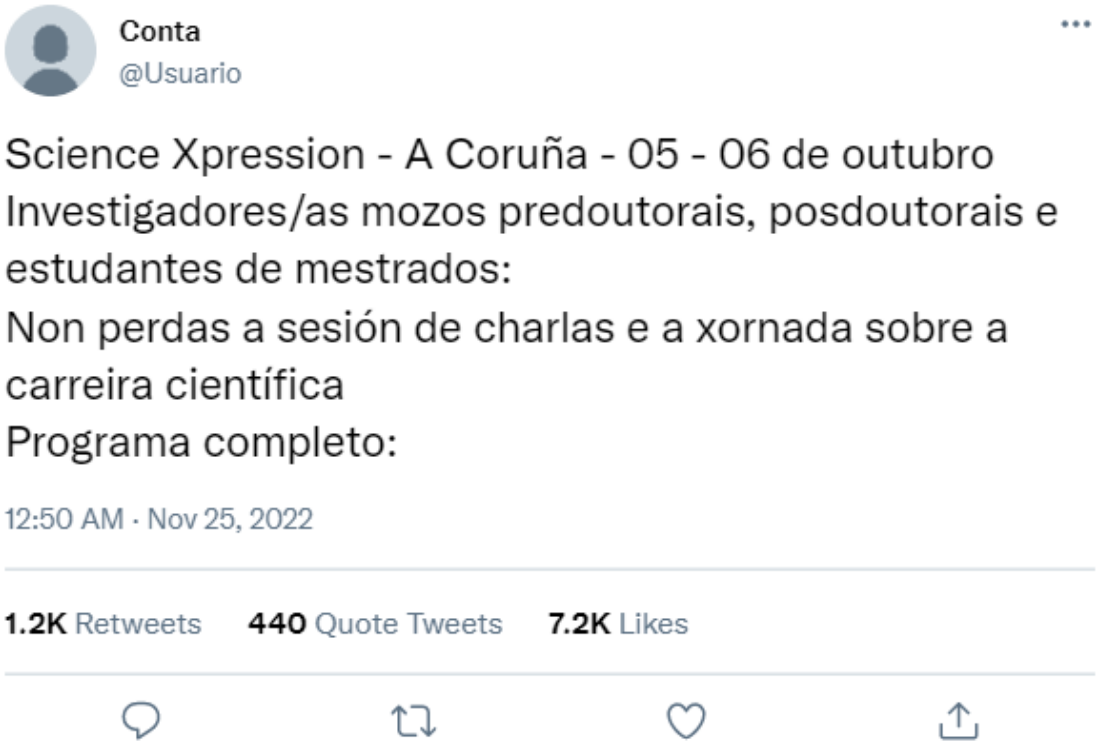


Ilustración 80. APC1.

Ilustración 81. APC2.

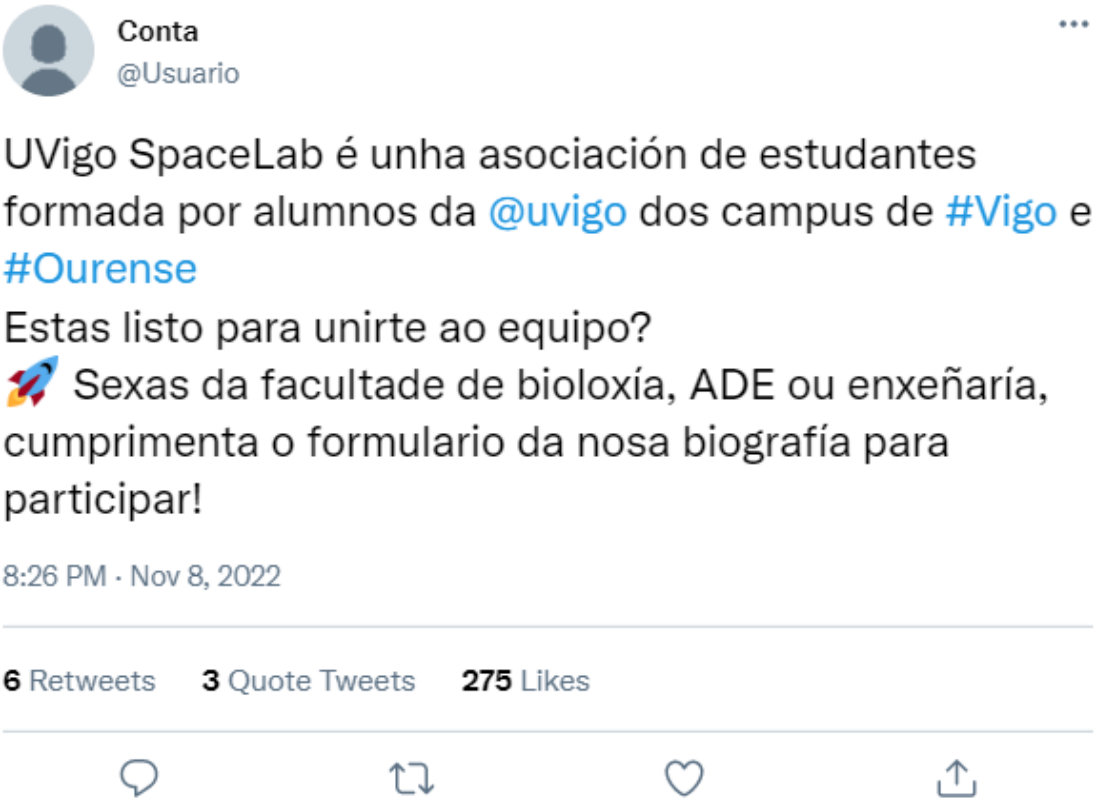


Ilustración 82. APC3.

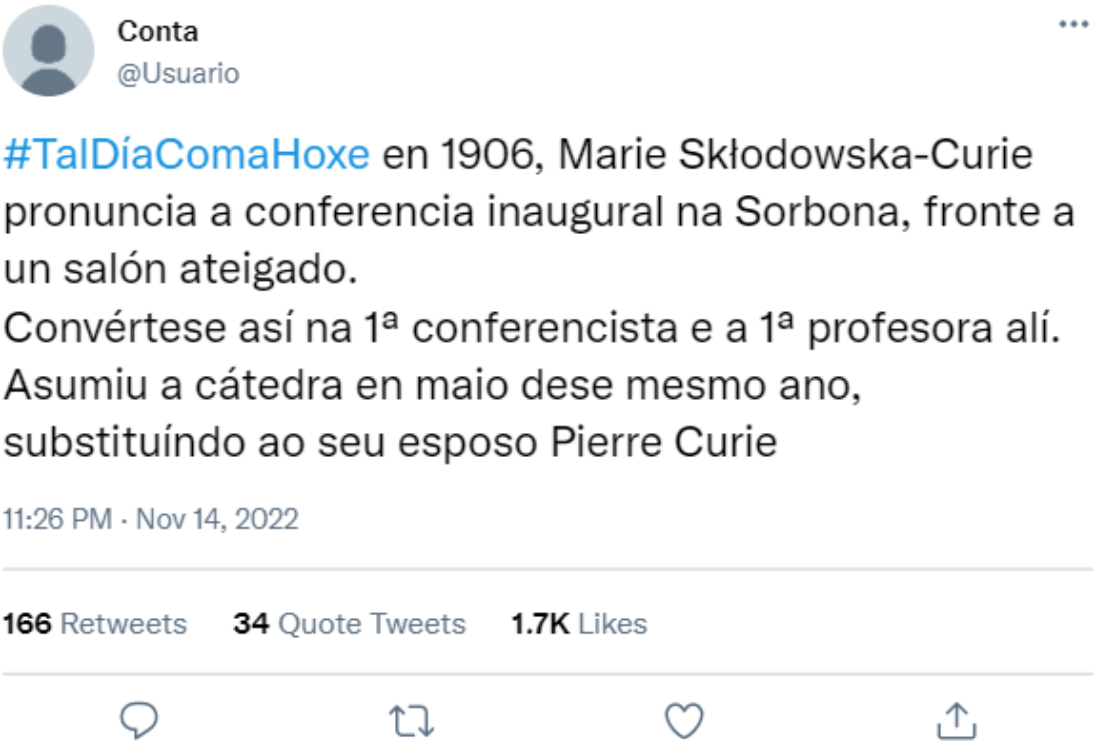


Ilustración 83. APL1.

Ilustración 84. APL2.

Ilustración 85. APL3.

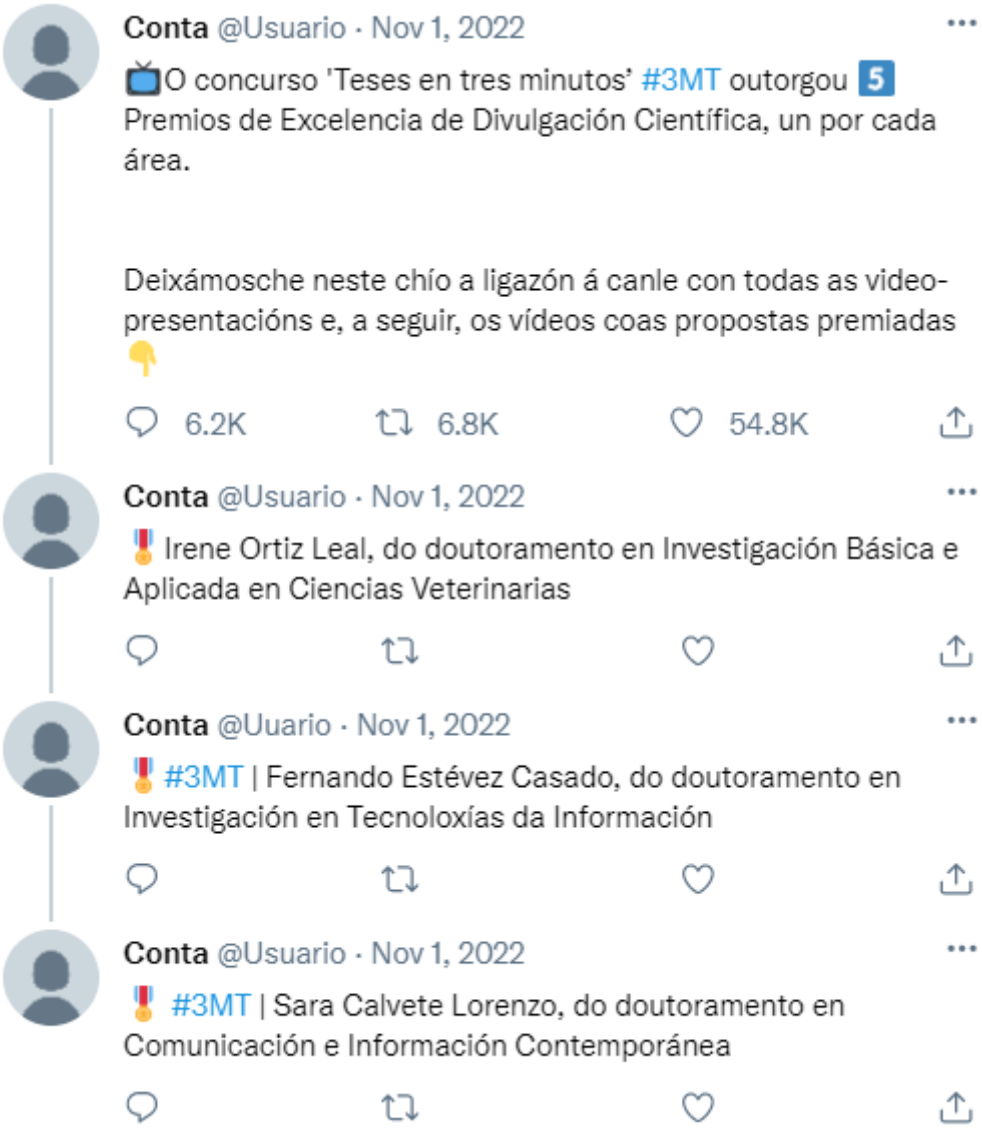


Ilustración 86. OFío1.

Ilustración 87. Ofío2.

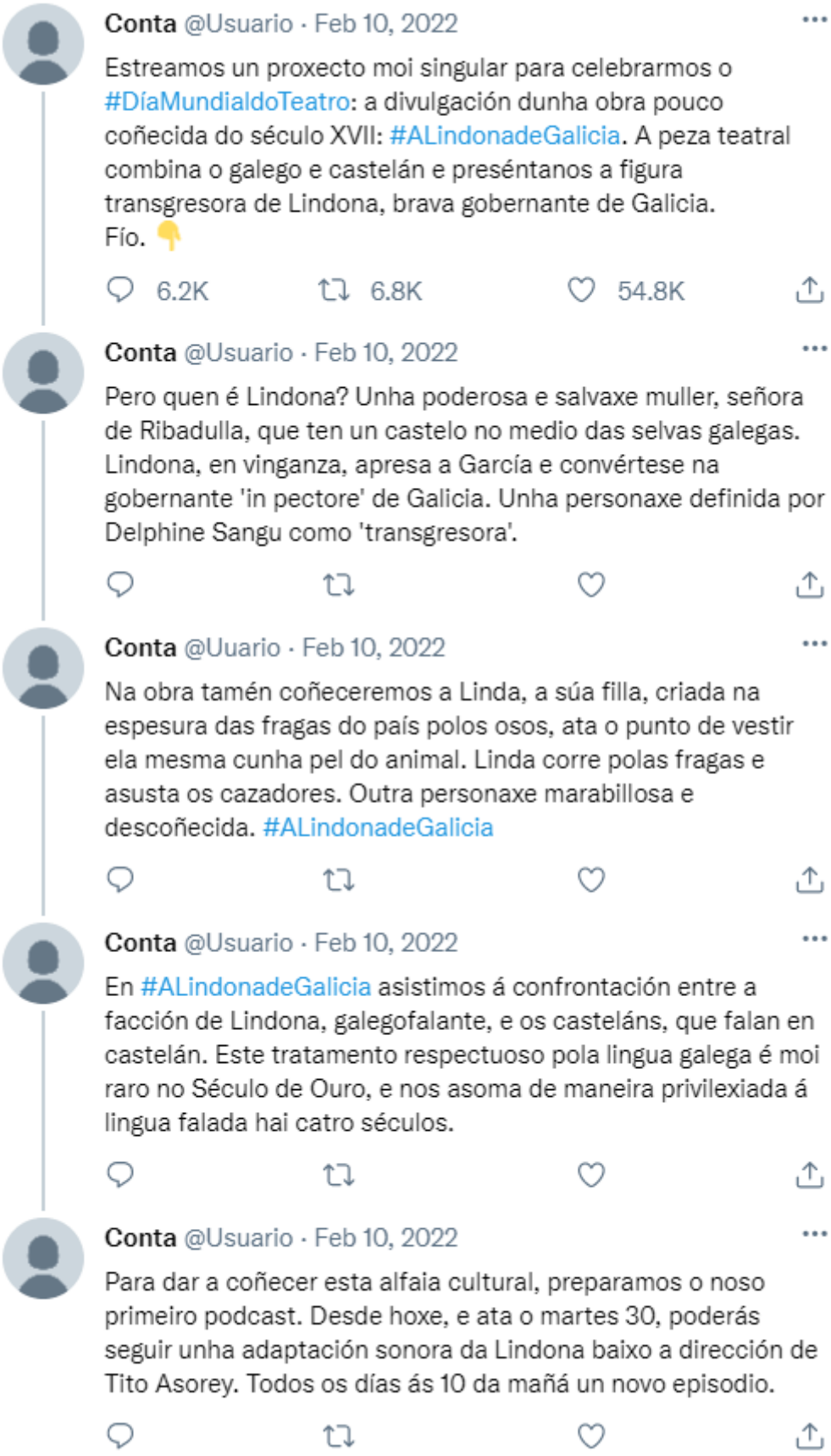


Ilustración 88. Ofío3.

Ilustración 89. OFC1.

Ilustración 90. OFC2.

Ilustración 91. OFC3.

Ilustración 92. OFL1.

Ilustración 93. OFL2.

Ilustración 94. OFL3.

Ilustración 95. OPC1.

Ilustración 96. OPC2.

Ilustración 97. OPC3.

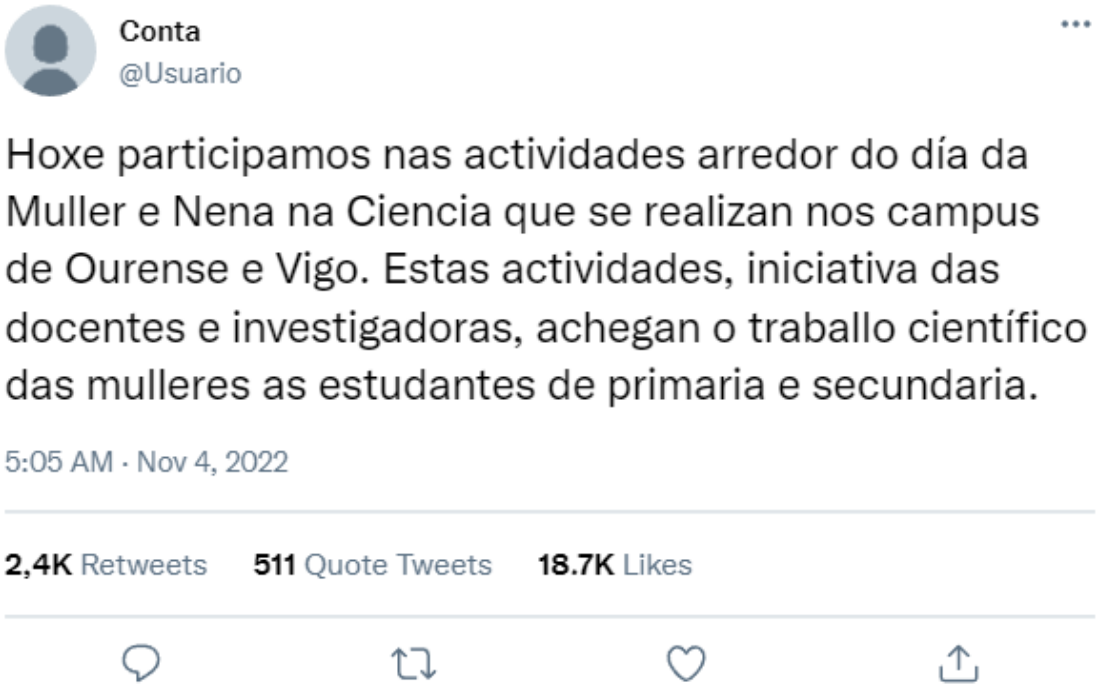
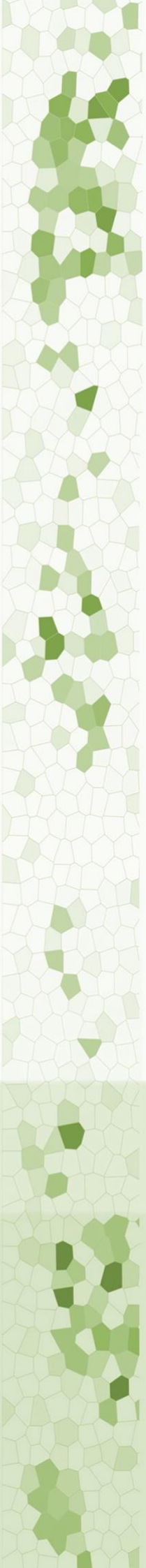


Ilustración 98. OPL1.

Ilustración 99. OPL2.

Ilustración 100. OPL3.



UAB
Universitat Autònoma
de Barcelona