Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 95-108 recibido 04-12-2025 revisado 18-01-2026 aceptado 12-02-2026

Desarrollo de modelos de TAN para las lenguas románicas de la península ibérica

Development of NMT Models for the Romance Languages of the Iberian Peninsula

Antoni Oliver,[1] **Sergi Álvarez-Vidal**[2]
[1]Universitat Oberta de Catalunya (UOC)
[2]Universitat Autònoma de Barcelona (UAB)
aoliverg@uoc.edu sergi.alvarez@uab.cat

Resumen: En este artículo se presentan los resultados finales del proyecto TAN-IBE, cuyo objetivo central ha sido diseñar sistemas de traducción automática neuronal (TAN) adaptados a las lenguas románicas de la península ibérica, con una especial atención a aquellas con menos recursos: el asturiano, el aragonés y el aranés. Este artículo se centra en presentar los corpus desarrollados y los modelos de traducción automática neuronal entrenados.
Palabras clave: traducción automática neuronal, lenguas románicas, lenguas con pocos recursos.

Abstract: This paper presents the final outcomes of the TAN-IBE project, whose central objective has been to develop Neural Machine Translation (NMT) systems specifically adapted for the Romance languages of the Iberian Peninsula, with a particular focus on the most under-resourced among them: Asturian, Aragonese, and Aranese. We present the developed corpora and the final neural machine translation systems trained.
Keywords: neural machine translation, romance languages, low-resource languages.

1 Introducción

En las dos últimas décadas, la traducción automática (TA) ha pasado de ser una tecnología experimental a ocupar un papel central en la comunicación multilingüe. La aparición de la TA neuronal (TAN) marcó un punto de inflexión respecto a los sistemas estadísticos tradicionales, al basarse en redes neuronales profundas capaces de modelar dependencias contextuales y producir traducciones más naturales (Koehn y Knowles, 2017). Sin embargo, los avances logrados por la TAN han sido desiguales: mientras que las lenguas con abundantes recursos y corpus extensos han experimentado mejoras significativas, las lenguas minoritarias o regionales siguen enfrentándose a una marginación digital persistente (Kornai, 2013; Guzmán et al., 2019).

Las denominadas *lenguas con pocos recursos* sufren la falta de datos paralelos y herramientas lingüísticas básicas, lo que dificulta su integración en los flujos de procesamiento automático. Este déficit de infraestructura tecnológica repercute directamente en su visibilidad en la red y en su uso en contextos institucionales o educativos, y contribuye a una brecha de acceso a contenidos, servicios y producción digital en la propia lengua (UNESCO, 2003; CE, 1992). En este sentido, la traducción automática puede desempeñar un papel estratégico como herramienta de inclusión digital.

En este contexto, el proyecto TAN-IBE tiene como objetivo principal desarrollar y evaluar sistemas de TAN que mejoren la traducción de lenguas románicas con escasos recursos del y hacia el español. A diferencia de otros enfoques centrados exclusivamente en la mejora de arquitecturas neuronales, el proyecto se orienta a la exploración práctica de estrategias que compensen la escasez de datos mediante la reutilización de recursos existentes y el aprovechamiento de la proximidad tipológica entre lenguas románicas, con énfasis en procedimientos reproducibles de recopilación, limpieza y evaluación de datos.

ISSN 1135-5948 DOI 10.26342/2026-76-7

El proyecto se apoya en tres pilares complementarios: (1) la creación y limpieza de corpus multilingües abiertos, (2) el entrenamiento y ajuste de modelos neuronales mediante técnicas de transferencia y *fine-tuning*, y (3) la evaluación empírica de los resultados obtenidos en términos de calidad de traducción y aplicabilidad en contextos reales.

El proyecto se articula en torno a una visión de ciencia abierta, en la que los datos, herramientas y modelos se publican bajo licencias libres, contribuyendo a los esfuerzos europeos de democratización tecnológica en el ámbito lingüístico (Burchell et al., 2025).

Los resultados de TAN-IBE confirman que la transferencia multilingüe y la generación controlada de datos sintéticos permiten obtener mejoras sustanciales incluso con recursos mínimos. Los experimentos realizados con asturiano, aragonés y aranés muestran que los modelos multilingües ajustados, como NLLB (Costa-jussà et al., 2022) o Marian (Junczys-Dowmunt et al., 2018a), pueden superar ampliamente a los sistemas basados en reglas y aproximarse al rendimiento de los modelos entrenados con lenguas de mayor volumen de datos.

Las estrategias aplicadas incluyen el uso de retrotraducción (Sennrich, Haddow, y Birch, 2016), el aprendizaje por transferencia (Zoph et al., 2016; Nguyen y Chiang, 2017), y la transferencia multilingüe en arquitecturas neuronales (Firat, Cho, y Bengio, 2016; Fan et al., 2021). Estos resultados indican que la proximidad lingüística puede compensar parcialmente la escasez de datos paralelos (Artetxe, Labaka, y Agirre, 2019; Lample et al., 2018).

Este artículo presenta una síntesis de los resultados finales del proyecto TAN-IBE, centrado en la traducción entre el español y tres lenguas románicas minoritarias: asturiano, aragonés y aranés.

2 Traducción automática neuronal para las lenguas románicas con pocos recursos

La traducción automática neuronal (TAN) ha redefinido el panorama del procesamiento del lenguaje natural en la última década. Desde los primeros modelos basados en atención (Bahdanau, Cho, y Bengio, 2015) hasta la arquitectura de transformadores (Vaswani et al., 2017), la calidad de las traducciones ha alcanzado niveles que permiten su integración en entornos profesionales y educativos. Sin embargo, ha ampliado la brecha entre lenguas con distinto grado de digitalización, ya que los modelos de última generación dependen de grandes volúmenes de datos paralelos y monolingües (Kornai, 2013; Caswell, Chelba, y Constant, 2020).

Las lenguas románicas constituyen un escenario idóneo para abordar esta problemática. Su parentesco morfosintáctico y léxico facilita la transferencia entre modelos, una ventaja especialmente relevante en el contexto iberorromance, donde el español, el portugués, el gallego y el catalán disponen de abundantes corpus paralelos, a diferencia de otras lenguas como el asturiano, el aragonés o el aranés. Esta relación de proximidad permite aplicar con éxito técnicas de *transfer learning*, mediante las cuales un modelo entrenado con una lengua con muchos recursos puede adaptarse a otra con pocos datos (Zoph et al., 2016; Nguyen y Chiang, 2017; Li et al., 2022).

En paralelo, los enfoques multilingües han demostrado su eficacia en la traducción de lenguas tipológicamente afines. Modelos como NLLB (Costa-jussà et al., 2022; Abdurakhmonova et al., 2024) o TowerInstruct (Alves et al., 2024) permiten aprovechar la representación compartida entre decenas de lenguas, lo que favorece la traducción *zero-shot* entre pares no vistos durante el entrenamiento (Fan et al., 2021; Xu et al., 2024).

Ante la escasez de corpus paralelos, las estrategias de generación de datos sintéticos se han consolidado como una solución central en la traducción automática para lenguas con pocos recursos. La retrotraducción (*backtranslation*) (Sennrich, Haddow, y Birch, 2016) permite aprovechar datos monolingües para ampliar el entrenamiento, mientras que la minería de corpus comparables en la web (Schwenk et al., 2021; Pan et al., 2021) facilita la identificación automática de segmentos potencialmente paralelos a gran escala. Estas técnicas suelen combinarse con métodos de filtrado basados en similitud semántica y *sentence embeddings* (Reimers y Gurevych, 2019), lo que mejora la calidad y coherencia de los datos generados.

En paralelo, diversas infraestructuras abiertas han contribuido a consolidar este ecosistema multilingüe. HPLT (Burchell et al., 2025) promueve la creación de grandes colecciones monolingües y herramientas de procesamiento para lenguas europeas infrarepre-

sentadas; OPUS-MT (Tiedemann y Thottingal, 2020) ofrece modelos neuronales entrenados sobre el repositorio OPUS y fomenta su reutilización abierta; y Bergamot (Huck y et al., 2020) ha impulsado la traducción automática integrada en navegadores mediante modelos eficientes ejecutables localmente. Estas iniciativas comparten un objetivo común: facilitar el acceso abierto a datos y modelos para reducir desigualdades tecnológicas entre lenguas.

No obstante, persisten retos estructurales específicos de las lenguas con pocos recursos, entre ellos la variación ortográfica, la inestabilidad normativa, la limitada disponibilidad de conjuntos de evaluación y la escasez de herramientas lingüísticas auxiliares. En el ámbito peninsular, el desafío consiste en articular estrategias técnicamente viables y metodológicamente reproducibles que permitan entrenar y evaluar modelos para lenguas románicas minoritarias en condiciones de fuerte asimetría de datos, sin depender exclusivamente del aumento cuantitativo del corpus. En este contexto, la TAN representa una tecnología de comunicación clave para mejorar la accesibilidad tecnológica de las lenguas románicas con pocos recursos.

3 Las lenguas románicas de la península ibérica

El espacio lingüístico de la península ibérica configura un mosaico de lenguas románicas que, a pesar de su origen común, presentan trayectorias sociolingüísticas y niveles de desarrollo tecnológico desiguales. Este conjunto incluye lenguas de amplia difusión, como el español y el portugués, y otras de ámbito regional o minoritario, como el catalán, el gallego, el asturiano, el aragonés y el aranés (variedad del occitano gascón). Su proximidad tipológica convierte al espacio peninsular en un entorno idóneo para investigar transferencia cruzada y enfoques multilingües en traducción automática neuronal.

El reconocimiento institucional y la presencia en la administración no garantizan por sí mismos una disponibilidad proporcional de recursos lingüísticos digitales. En la práctica, la cantidad de corpus paralelos y monolingües depende de factores como la producción textual acumulada, la digitalización histórica y la incorporación temprana a infraestructuras de datos abiertas. OPUS (Tiedemann, 2012), Common Crawl (Wenzek et al., 2020) o Paracrawl (Bañón et al., 2020) incluyen millones de segmentos para lenguas con amplia presencia editorial, mediática y web, mientras que otras variedades presentan una disponibilidad mucho más reducida. El asturiano y el aragonés muestran una escasez estructural de datos que condiciona el entrenamiento de modelos neuronales robustos. El aranés constituye un caso singular: pese a su cooficialidad en Cataluña, su reducido número de hablantes y su limitada masa textual digital hacen que su representación en grandes repositorios siga siendo modesta.

El catalán destaca inicialmente como la lengua minoritaria con mayor infraestructura tecnológica, gracias a herramientas de código abierto como Apertium (Forcada et al., 2011), OpusMT (Tiedemann y Thottingal, 2020) o el proyecto Aina (Gonzalez-Agirre et al., 2024). El gallego muestra un desarrollo intermedio, reforzado por corpus paralelos y proyectos recientes (de Dios-Flores et al., 2022). Como se aprecia en la Tabla 1, la desigual disponibilidad de datos condiciona las estrategias de modelado y obliga a recurrir a técnicas como la retrotraducción (Sennrich, Haddow, y Birch, 2016) o la transferencia entre lenguas próximas (Zoph et al., 2016; Nguyen y Chiang, 2017).

4 El proyecto TAN-IBE

El proyecto TAN-IBE se ha organizado en las siguientes tareas:

4.1 Recopilación y limpieza de los corpus existentes

En el proyecto se han recopilado tanto corpus monolingües como paralelos ya existentes. Estos corpus a menudo contienen mucho ruido, por ejemplo, segmentos que no están en la lengua deseada, y, en el caso de corpus paralelos, pares de segmentos que no son equivalentes de traducción. Durante el proyecto se ha mejorado una herramienta propia de *rescoring* de corpus[1] (Oliver y Álvarez, 2023). También se ha desarrollado un modelo de detección de lenguaje capaz de detectar con mayor precisión el asturiano y el aragonés, así como distinguir entre aranés y el resto de variantes del occitano.

Los corpus monolingües del proyecto HPLT han sido sometidos a un proceso de

[1] `https://github.com/mtuoc/MTUOC-PCorpus-rescorer`

Lengua	Estatus institucional	Hablantes	Segmentos OPUS
Portugués	Oficial estatal e internacional	11 millones	306,4 M
Catalán	Cooficial en 3 CCAA y oficial en Andorra	10 millones	99,0 M
Gallego	Cooficial en Galicia	2,5 millones	36,3 M
Asturiano	Reconocido, no oficial	0,2 millones	6,9 M
Aragonés	Reconocido, no oficial	0,03 millones	62,2 K
Aranés (occitano)	Cooficial en Cataluña	0,006 millones	1,1 M

Tabla 1: Situación institucional y disponibilidad estimada de corpus paralelos con traducción al español en OPUS para las lenguas románicas de la península ibérica.

limpieza que incluye la verificación automática de la lengua de cada segmento mediante nuestro modelo de detección de lengua. Se han eliminado aquellos casos en los que la lengua objetivo presentaba un índice de confianza inferior a 0,75. También se ha utilizado el corpus PILAR (Sánchez-Martínez et al., 2024), que incorpora textos en asturiano, aragonés y aranés. La tabla 2 muestra el tamaño final de los corpus monolingües disponibles para las lenguas de interés.

En cuanto a los corpus paralelos, se han utilizado los conjuntos disponibles en NLLB (Costa-jussà et al., 2022) y Aina (Gonzalez-Agirre et al., 2024). En este caso, el procedimiento de limpieza se ha aplicado a ambas lenguas de cada par (origen y destino), manteniendo únicamente aquellos segmentos cuya detección lingüística alcanzaba una confianza mínima de 0,75. Además, se ha calculado un índice de similitud semántica mediante un modelo multilingüe de *sentence embeddings*. Los pares con una distancia coseno inferior a 0,75 han sido descartados, con el fin de garantizar un alineamiento de calidad razonable. La tabla 3 recoge el tamaño de los corpus paralelos resultantes tras estos filtros. Cabe señalar que el corpus Aina combina datos paralelos reales, segmentos retrotraducidos y corpus completamente sintético. Por este motivo, en nuestros experimentos todos los segmentos procedentes de Aina se han tratado como datos sintéticos durante el entrenamiento de los modelos.

4.2 Compilación, preprocesado y limpieza de nuevos corpus

Durante el proyecto se han recopilado nuevos corpus para el asturiano, aragonés y aranés. Se han obtenido documentos digitalizados, se han descargado páginas web y se ha utilizado la Wikipedia (*dumps* de 1 de marzo de 2025). No todos los corpus se pueden distribuir, ya que algunos documentos se han cedido solo para investigación y para entrenar los motores y otros documentos descargados de Internet no tenían una licencia clara. En esta sección detallamos únicamente los datos de los corpus que se distribuyen como resultados del proyecto.

Para el aranés, no existe una edición propia de Wikipedia. Conviene recordar que el aranés es una de las variantes del occitano. En la Wikipedia occitana, los artículos incluyen una etiqueta que indica en qué variante están redactados. Sin embargo, el número de artículos escritos en aranés es muy reducido. Por este motivo, hemos compilado por separado un corpus formado por los textos de los artículos escritos en las demás variantes del occitano.

El primer paso del preprocesamiento consiste en la segmentación del texto, es decir, en su división en unidades mínimas de análisis. En el contexto de los corpus paralelos, un segmento suele corresponder a una oración gramatical completa; no obstante, dependiendo del objetivo analítico, puede también consistir en palabras individuales, grupos léxicos o expresiones numéricas. En términos operativos, el segmento constituye la unidad mínima con entidad funcional o semántica relevante para el tipo de procesamiento posterior. La segmentación se ha llevado a cabo mediante un conjunto de reglas definidas en formato SRX (*Segmentation Rules eXchange*), estándar para la delimitación de unidades textuales en entornos de traducción y procesamiento lingüístico. En el marco del proyecto, se han desarrollado reglas específicas adaptadas al asturiano, aragonés, aranés y occitano.

Una vez segmentados los corpus, se ha procedido a detectar segmentos paralelos. Dado que los artículos de la Wikipedia en dos lenguas pueden ser artículos independientes, se ha llevado a cabo la tarea conocida

Corpus	Asturiano	Aragonés	Occitano	Aranés
HPLT	3.375.722	-	1.540.352	101.274
PILAR	38.839	84.697	-	312.477
TOTAL uniq.	3.414.315	84.697	1.540.352	413.063

Tabla 2: Tamaño en segmentos de los corpus monolingües existentes.

Corpus	Asturiano	Aragonés	Occitano	Aranés
NLLB	504.532	-	108.834	-
Aina	206.636	47.506	-	2.199.919

Tabla 3: Tamaño en segmentos de los corpus paralelos existentes.

como *bitext mining* (Sharami, Sterionov, y Spronck, 2021) y que hemos implementado en la herramienta MTUOC-aligner.[2]

Una de las principales dificultades que nos hemos encontrado durante la recopilación de los corpus ha sido la existencia de diversas normas ortográficas para el aragonés y la aparición, una vez empezado el proyecto, de una nueva normativa de la Academia Aragonesa de la Lengua. Los corpus que distribuimos como resultados del proyecto cumplen esta nueva normativa ortográfica, que no es usada de forma unánime a pesar de ser la oficial.

Además de los corpus monolingües y paralelos se han desarrollado corpus retrotraducidos (*backtranslated*) (Edunov et al., 2018) y sintéticos. Un corpus retrotraducido de la lengua A a la lengua B se construye traduciendo automáticamente a la lengua A los segmentos monolingües disponibles en la lengua B. Por el contrario, un corpus sintético de la lengua A a la lengua B se obtiene traduciendo automáticamente a la lengua B los segmentos monolingües originales en la lengua A. Para la creación de estos recursos se ha utilizado principalmente Apertium, excepto para asturiano-español, que no está disponible. En este caso, se ha utilizado un motor neuronal preliminar entrenado para este propósito. Los corpus sintéticos se han obtenido a partir de la parte española del corpus NTEU español-inglés (Bié et al., 2020). Dado que Apertium es capaz de identificar las palabras desconocidas, se han eliminado todos los pares de segmentos que contenían palabras desconocidas en la traducción. Como consecuencia, el tamaño final de los corpus sintéticos varía entre las distintas combinaciones lingüísticas, ya que el número de palabras desconocidas difiere según la lengua implicada.

[2] `https://github.com/mtuoc/MTUOC-aligner`

4.3 Entrenamiento de sistemas TAN con diversas técnicas

En el proyecto TAN-IBE se han entrenado motores de traducción entre el español y el portugués, gallego, catalán, asturiano, aragonés y aranés, en ambas direcciones. Además, se ha desarrollado un modelo multilingüe capaz de traducir del español a todas las demás lenguas del proyecto. A lo largo del desarrollo se han probado diversas técnicas y configuraciones.

En la sección 5 presentamos los modelos bilingües finales entre el español y el asturiano, aragonés y aranés, junto con sus resultados de evaluación. Estos modelos se entrenaron utilizando los corpus existentes y los distribuidos por el proyecto, sin recurrir a corpus recopilados específicamente durante su desarrollo con licencias o cesiones de derechos limitadas a usos de investigación. De este modo, garantizamos que los resultados presentados en este artículo sean plenamente reproducibles.

En cada subapartado se detalla la composición del corpus de entrenamiento para cada par de lenguas, así como los pesos utilizados. Para todos los pares de lenguas se ha utilizado como corpus de validación un fragmento dev (997 segmentos) del corpus Flores+ (Pérez-Ortiz et al., 2024).

Para preprocesar los corpus se ha utilizado sentencepiece (Kudo y Richardson, 2018) empleando una estrategia de *Byte Pair Encoding* (BPE) y con un tamaño de vocabulario de 64.000 subpalabras. El procesamiento se ha realizado conjuntamente con las dos lenguas y se ha obtenido un vocabulario común.

Para el entrenamiento se ha utilizado Marian (Junczys-Dowmunt et al., 2018b) con una configuración de tipo *Transformer*. En cuanto a la arquitectura, utiliza un dimensionamiento de *embedding* de 512 y el bloque *Feed-Forward Network* (FFN) en el Transformer tiene una dimensión interna de 2.048, con

Corpus	Tipo	Lenguas	Segmentos
wikipedia-ast	monolingüe	asturiano	2.233.988
wikipedia-spa-ast	paralelo	español-asturiano	1.539.239
wikipedia-backtranslated-spa-ast	paralelo backtranslated	español-asturiano	2.333.847
NTEU-synthetic-spa-ast	paralelo sintético	español-asturiano	77.220.441
wikipedia-AAL-arg	monolingüe	aragonés	349.069
wikipedia-AAL-spa-arg	paralelo	español-aragonés	6.051
wikipedia-AAL-backtranslated-spa-arg	paralelo backtranslated	español-aragonés	52.810
NTEU-AAL-synthetic-AAL-spa-arg	paralelo sintético	español-aragonés	5.019.629
wikipedia-oci_aran	monolingüe	aranés	1.123
wikipedia-spa-oci_aran	paralelo	español-occitano	17
wikipedia-backtranslated-spa-oci_aran	paralelo backtranslated	español-aranés	1.123
NTEU-synthetic-spa-oci_aran	paralelo sintético	español-aranés	5.087.903
wikipedia-oci	monolingüe	occitano	229.799
wikipedia-spa-oci	paralelo	español-occitano	1.426
wikipedia-backtranslated-spa-oci	paralelo backtranslated	español-occitano	229.522

Tabla 4: Corpus distribuidos por el proyecto.

8 cabezales de atención. Para la regularización, se aplica un *dropout* de 0,2 al Transformer. El entrenamiento se optimiza con el algoritmo Adam (Kingma, 2014) utilizando una tasa de aprendizaje inicial de 0,0001 y el tamaño de los *mini-batches* es de 64. La validación se realiza cada 10.000 *updates* utilizando la entropía cruzada y la métrica BLEU y se incorpora un mecanismo de parada temprana de 10. Recordemos que en el entrenamiento se han utilizado diferentes pesos según el corpus de procedencia de los segmentos, asignando 1 a los corpus paralelos, 0.75 a los retrotraducidos y 0.5 a los sintéticos.

Para la dirección español - aranés se ha entrenado un sistema multilingüe capaz de traducir tanto del español al aranés como al occitano. Se han utilizado marcas de dirección en la parte española de los corpus: <2oci_aran>para los segmentos con equivalentes en aranés y <2oci>para los segmentos con equivalentes en occitano.

4.4 Evaluación de los modelos entrenados y comparación con otros modelos existentes

Los modelos entrenados se comparan con una serie de modelos y en la mayoría de casos se considera Apertium como *baseline* (excepto para la dirección asturiano-español, ya que Apertium no está disponible). Apertium es un sistema de TA basado en reglas con una metodología de transferencia sintáctica superficial. Los modelos con los que comparamos son los modelos TAN bilingües Aina (Sant et al., 2024), el modelo multilingüe NLLB (Costa-jussà et al., 2022) y SalamandraTA (Garcia Gilabert et al., 2025), así como SalamandraTA-Aranese, una versión de SalamandraTA ajustada específicamente para esta lengua. Hay que tener en cuenta que no todos estos modelos están disponibles para todos los pares de lenguas analizados. Se ha seleccionado Apertium como modelo base para la comparación porque es un sistema ampliamente utilizado para estos pares de lenguas y, al distribuirse bajo licencia libre, ha sido la base para otros sistemas derivados. El resto de modelos de la comparación se han escogido porque son modelos libremente disponibles y que obtienen resultados competitivos. Los modelos del proyecto Aina participaron en la *Shared Task Machine Translation for Romance Languages of the Iberian Peninsula* (Sánchez-Martínez et al., 2024), organizada en el marco de la WMT24.

Para la evaluación de los sistemas, hemos utilizado únicamente las métricas automáticas disponibles en SacreBLEU[3] (Post, 2018). Cada una de estas métricas valora diversos aspectos y sus resultados son complementarios:

- BLEU (*Bilingual Evaluation Understudy*) (Papineni et al., 2002): mide la precisión de n-gramas, siendo el estándar de la industria para evaluar la fidelidad léxica frente a la referencia.
- chrF++ (Popović, 2016): se basa en la coincidencia de caracteres y n-gramas de palabras. Es especialmente útil para lenguas con morfología rica (como las romances tratadas aquí), ya que correlaciona mejor con el juicio humano al penalizar menos los errores de flexión.

[3] `https://github.com/mjpost/sacrebleu`

- TER (*Translation Edit Rate*) (Snover et al., 2006): mide el esfuerzo de edición necesario para transformar la salida del sistema en la referencia. A diferencia de las anteriores, aporta una perspectiva sobre la eficiencia y la utilidad práctica del texto generado.

Se han realizado test de significancia estadística comparando cada sistema con el de referencia mediante la técnica de *paired bootstrap resampling* (con 1.000 iteraciones), integrada en la herramienta SacreBLEU. Utilizando un umbral de significancia de 0,05, los resultados con un $p < 0,05$ permiten rechazar la hipótesis nula, indicando que las diferencias observadas no son fruto del azar. Cabe destacar que este test no determina por sí mismo si un sistema es mejor, sino que confirma la replicabilidad de la diferencia entre los modelos evaluados y el modelo de referencia.

No utilizamos métricas neuronales como COMET[4] (Rei et al., 2020) o BLEURT[5] (Sellam, Das, y Parikh, 2020), ya que el asturiano, el aragonés y el aranés no están bien representados en los modelos preentrenados disponibles. Como corpus de evaluación se ha utilizado el fragmento devtest (1.012 segmentos) del corpus Flores+.

5 Modelos entrenados y evaluación

5.1 Asturiano

Para entrenar los modelos del español y asturiano en ambas direcciones, se han utilizado los corpus mostrados en la tabla 5. La tabla muestra tanto el número de segmentos como el peso que se le ha dado a cada corpus durante el entrenamiento.

5.1.1 Modelo español - asturiano

La tabla 6 presenta la evaluación de sistemas de traducción automática para la combinación español-asturiano, utilizando las métricas BLEU, chrF2 y TER, comparados con el sistema de referencia (Apertium). Como podemos observar, SalamandraTA, que es un LLM afinado para traducción, es el que consigue mejores resultados en todas las métricas. De entre los modelos de traducción neuronal, el sistema Aina muestra el mejor rendimiento en las tres métricas y nuestro modelo queda en segunda posición. Nuestro sistema supera a Apertium en todas las métricas con diferencias estadísticamente significativas. En cambio, el modelo NLLB analizado obtiene peores resultados de BLEU y chrF2 que Apertium, pero mejores en TER.

5.1.2 Modelo asturiano - español

La tabla 7 muestra la evaluación de dos sistemas de traducción automática, Salamandra TA 7b instruct y nuestro modelo (Marian TAN-IBE), para la dirección asturiano-español, comparándolos con el sistema de referencia, que en este caso es nllb-200-distilled-600M, ya que el sistema Apertium no está disponible para esta dirección. Ambos sistemas evaluados superan al sistema de referencia en todas las métricas. Salamandra TA 7b instruct obtiene la mejor puntuación BLEU por un margen significativo, mientras que Marian TAN-IBE logra el mejor resultado en chrF2 y TER.

5.2 Aragonés

Para entrenar el modelo español - aragonés, en ambas direcciones, se han utilizado los corpus indicados en la tabla 8.

5.2.1 Modelo español - aragonés

En la tabla 9 podemos observar los valores de evaluación de los sistemas español - aragonés. El modelo Aina muestra un rendimiento muy similar al sistema de referencia Apertium, con una ligera disminución en BLEU (60,1 frente a 61,1) que es estadísticamente significativa, pero con puntuaciones de chrF2 (79,4 frente a 79,3) y TER (27,3 frente a 27,2) cuyas diferencias no son estadísticamente significativas. Nuestro modelo Marian TAN-IBE también obtiene resultados muy cercanos a Apertium, aunque consistentemente peores con diferencias estadísticamente significativas en todas las métricas. Por otro lado, Salamandra TA 7b instruct presenta un rendimiento considerablemente inferior en todas las métricas, indicando que para este par de lenguas es el menos efectivo de los sistemas probados y que estas diferencias son altamente significativas. En conclusión, Apertium sigue siendo el sistema con mejor rendimiento, con Aina y Marian TAN-IBE ofreciendo un rendimiento muy cercano.

5.2.2 Modelo aragonés - español

En la tabla 10 se puede observar la evaluación de los sistemas de traducción automática para la dirección aragonés - español, to-

[4] https://unbabel.github.io/COMET/html/index.html

[5] https://github.com/google-research/bleurt

Corpus	Segmentos	Peso
Wikipedia	1.539.239	1
NLLB rescored	504.532	1
Wikipedia backtranslated	2.333.536	0.75
HPLT backtranslated	3.375.707	0.75
PILAR backtranslated	38.838	0.75
NTEU sintético	7.220.441	0.5
Aina	206.636	0.5

Tabla 5: Tamaño y peso de los corpus utilizados para entrenar el motor español - asturiano.

System	BLEU ($\mu\pm$ 95 % CI)	chrF2 ($\mu\pm$ 95 % CI)	TER ($\mu\pm$ 95 % CI)
Baseline: Apertium	17.0 (± 0.7)	50.8 (± 0.6)	80.4 (± 1.2)
NLLB-200-distilled-600M	15.6 (± 0.7) (p = 0.0010)*	49.8 (± 0.6) (p = 0.0010)*	78.3 (± 1.9) (p = 0.0050)*
Aina	18.2 (± 0.7) (p = 0.0010)*	52.1 (± 0.6) (p = 0.0010)*	74.1 (± 1.2) (p = 0.0010)*
Marian TAN-IBE	17.8 (± 0.7) (p = 0.0010)*	51.3 (± 0.6) (p = 0.0010)*	76.0 (± 1.1) (p = 0.0010)*
SalamandraTA	**21.0** (± 0.9) (p = 0.0010)*	**53.2** (± 0.7) (p = 0.0010)*	**70.5** (± 2.0) (p = 0.0010)*

Tabla 6: Evaluación de los motores español - asturiano.

System	BLEU (μ ± 95 % CI)	chrF2 (μ ± 95 % CI)	TER (μ ± 95 % CI)
Baseline: nllb-200-distilled-600M	23.6 (23.6 ± 0.8)	62.7 (62.7 ± 0.6)	58.3 (58.3 ± 0.9)
SalamandraTA 7b instruct	**34.8** (34.8 ± 1.8) (p = 0.0010)*	64.8 (64.8 ± 1.1) (p = 0.0010)*	51.8 (51.8 ± 1.7) (p = 0.0010)*
Marian TAN-IBE	29.4 (29.4 ± 0.9) (p = 0.0010)*	**68.6** (68.6 ± 0.5) (p = 0.0010)*	**51.1** (51.1 ± 0.9) (p = 0.0010)*

Tabla 7: Evaluación de los motores asturiano - español.

mando como referencia Apertium. Nuestro modelo Marian TAN-IBE supera a Apertium en todas las métricas, pero SalamandraTA es el que presenta unos mejores resultado. Este comportamiento confirma la tendencia observada en otros experimentos, que los LLM afinados para la traducción ofrecen resultados notables para las direcciones entre una lengua con pocos recursos y una con muchos recursos.

5.3 Aranés

Para entrenar el modelo español - aranés, en ambas direcciones, se han utilizado los corpus mostrados en la tabla 11. Si nos fijamos en la tabla, el número de segmentos disponibles para español - aranés es significativamente inferior a los disponibles para español - asturiano y español - aragonés. Por este motivo se ha decidido también utilizar los corpus español - occitano que se muestran en la tabla 12.

5.3.1 Modelo español - aranés

La tabla 13 presenta la evaluación de los modelos de traducción automática para la dirección español - aranés, utilizando el sistema Apertium como referencia. Como se puede observar, los mejores resultados de BLEU y TER los consigue nuestro motor, mientras que Aina obtiene los mejores resultados de chrF2. Para esta dirección de traducción tanto el LLM SalamandraTA general, como el afinado específicamente para el aranés, obtienen resultados peores que nuestro motor.

5.3.2 Modelo aranés - español

En la tabla 14 se muestran los resultados de evaluación de los modelos para la dirección aranés - español. De nuevo, para una traducción de una lengua con pocos recursos a una lengua con muchos recursos, los LLM afinados para traducción obtienen los mejores resultados. Destaca notablemente el modelo afinado para el aranés, que obtiene una mejora en BLEU de 1.1 puntos respecto al SalamandraTA genérico y de 11.5 puntos respecto a nuestro sistema neuronal.

6 Limitaciones de este estudio

Como conjunto de evaluación se han utilizado los fragmentos devtest de Flores+. Las versiones aragonesa y aranesa de este corpus se han desarrollado mediante traducción automática de la versión española más

Corpus	Segmentos	Peso
Wikipedia	6.051	1
Wikipedia backtranslated	52.810	0.75
PILAR backtranslated	84.697	0.75
NTEU sintético	5.019.629	0.5
Aina	47.500	0.5

Tabla 8: Tamaño y peso de los corpus utilizados para entrenar el motor español - aragonés.

System	BLEU ($\mu\pm$ 95 % CI)	chrF2 ($\mu\pm$ 95 % CI)	TER ($\mu\pm$ 95 % CI)
Baseline: Apertium	**61.1** (± 1.2)	79.3 (± 0.8)	**27.2** (± 1.1)
Aina	60.1 (± 1.2) (p = 0.0030)*	**79.4** (± 0.7) (p = 0.1578)	27.3 (± 1.1) (p = 0.1508)
Marian TAN-IBE	60.3 (± 1.3) (p = 0.0010)*	78.7 (± 0.8) (p = 0.0010)*	28.0 (± 1.2) (p = 0.0010)*
SalamandraTA	45.2 (± 1.4) (p = 0.0010)*	71.6 (± 0.8) (p = 0.0010)*	38.5 (± 1.8) (p = 0.0010)*

Tabla 9: Evaluación de los motores español - aragonés.

System	BLEU ($\mu\pm$ 95 % CI)	chrF2 ($\mu\pm$ 95 % CI)	TER ($\mu\pm$ 95 % CI)
Baseline: Apertium	61.6 (± 1.2)	77.9 (± 0.8)	26.6 (± 1.1)
Marian TAN-IBE	63.9 (± 1.2) (p = 0.0010)*	78.6 (± 0.8) (p = 0.0010)*	25.4 (± 1.1) (p = 0.0010)*
SalamandraTA	**67.1** (± 1.2) (p = 0.0010)*	**80.0** (± 0.8) (p = 0.0010)*	**23.5** (± 1.0) (p = 0.0010)*

Tabla 10: Evaluación de los motores aragonés - español.

una postedición humana llevada a cabo con traductores profesionales (Pérez-Ortiz et al., 2024). Esto da ventaja al sistema Apertium y puede enmascarar las mejoras conseguida por el resto de sistemas.

No se ha llevado a cabo una evaluación humana ni un estudio de la usabilidad práctica de los motores desarrollados.

7 Conclusiones y trabajo futuro

El proyecto TAN-IBE ha logrado demostrar que es posible desarrollar sistemas de traducción automática neuronal (TAN) competitivos para lenguas románicas de la península ibérica con pocos recursos, como el asturiano, el aragonés y el aranés, ayudando a luchar contra la histórica marginación digital que sufren estas lenguas. Los resultados confirman que la generación controlada de datos sintéticos (retrotraducción y uso de corpus sintéticos) es efectiva para compensar la escasez de corpus paralelos. En la mayoría de los pares de traducción evaluados, los modelos ajustados (Marian TAN-IBE y Aina) superan o igualan a los sistemas tradicionales basados en reglas (Apertium).

Durante la realización del proyecto se ha producido un paso progresivo del uso de modelos TAN a LLM afinados para la traducción. En este trabajo se ha comprobado que el modelo SalamandraTA obtiene resultados muy competitivos para estas lenguas, especialmente en la dirección de la lengua con menos recursos a la lengua con más recursos. Una vez finalizado el proyecto TAN-IBE en agosto de 2025, se ha dado inicio a un nuevo proyecto de investigación que explorará más a fondo el uso de LLMs afinados para la traducción para estas lenguas con pocos recursos y añadiendo una lengua más con extremadamente pocos recursos: el eonaviego. Se trata del proyecto LLMTrad-IBE (*Grandes modelos del lenguaje para la traducción de lenguas románicas de la península Ibérica con pocos recursos*), subproyecto del proyecto coordinado AI-TraLow (*Traducción para lenguas y culturas con pocos recursos guiada por la inteligencia artificial*).

En conclusión, el proyecto TAN-IBE no solo contribuye con modelos de traducción de código abierto con un rendimiento competitivo para las lenguas románicas de la península ibérica con menos recursos, sino que también ofrece herramientas y recursos que facilitan investigaciones futuras. El enfoque del proyecto, basado en la ciencia abierta y la reutilización de recursos, sienta un precedente crucial para el desarrollo de infraestructuras lingüísticas equitativas. Los resultados finales demuestran que la brecha digital en la tra-

Corpus	Segmentos	Peso
Wikipedia	17	1
Wikipedia backtranslated	1.123	0.75
PILAR backtranslated	312.477	0.75
HPLT backtranslated	214.291	0.75
NTEU sintético	5.087.903	0.5
Aina	219.919	0.5

Tabla 11: Tamaño y peso de los corpus español - aranés utilizados para entrenar el motor español - aranés.

Corpus	Segmentos	Peso
Wikipedia	1.426	1
NLLB	108.834	1
Wikipedia backtranslated	229.522	0.75
HPLT backtranslated	535.531	0.75

Tabla 12: Tamaño y peso de los corpus español - occitano utilizados para entrenar el motor español - aranés.

System	BLEU ($\mu\pm$ 95 % CI)	chrF2 ($\mu\pm$ 95 % CI)	TER ($\mu\pm$ 95 % CI)
Baseline: Apertium	50.8 (± 1.2)	73.4 (± 0.7)	36.2 (± 1.2)
Aina	49.3 (± 1.2) (p = 0.0010)*	**74.3** (± 0.8) (p = 0.0010)*	38.3 (± 1.2) (p = 0.0010)*
Marian TAN-IBE	**51.7** (± 1.2) (p = 0.0010)*	73.8 (± 0.7) (p = 0.0010)*	**35.7** (± 1.2) (p = 0.0010)*
SalamandraTA	46.7 (± 1.1) (p = 0.0010)*	70.3 (± 0.8) (p = 0.0010)*	39.5 (± 1.2) (p = 0.0010)*
SalamandraTA-Aranese	50.9 (± 1.2) (p = 0.3057)	73.0 (± 0.7) (p = 0.0450)*	36.7 (± 1.2) (p = 0.0549)

Tabla 13: Evaluación de los motores español - aranés.

System	BLEU ($\mu\pm$ 95 % CI)	chrF2 ($\mu\pm$ 95 % CI)	TER ($\mu\pm$ 95 % CI)
Baseline: Apertium	46.0 (± 1.1)	69.3 (± 0.8)	38.4 (± 1.1)
Marian	50.0 (± 1.1) (p = 0.0010)*	70.7 (± 0.8) (p = 0.0010)*	38.9 (± 1.1) (p = 0.0090)*
SalamandraTA	60.4 (± 1.2) (p = 0.0010)*	76.1 (± 0.8) (p = 0.0010)*	29.0 (± 1.1) (p = 0.0010)*
SalamandraTA-Aranese	**61.5** (± 1.3) (p = 0.0010)*	**76.7** (± 0.8) (p = 0.0010)*	**28.4** (± 1.1) (p = 0.0010)*

Tabla 14: Evaluación de los motores aranés - español.

ducción automática para lenguas con pocos recursos se puede reducir de manera efectiva, promoviendo la inclusión digital y la sostenibilidad tecnológica de estas lenguas.

Todas las herramientas, corpus y modelos generados durante el proyecto TAN-IBE se pueden obtener desde la página web del proyecto.[6]

Agradecimientos

Este trabajo ha sido financiado por los proyectos *TAN-IBE: Traducción automática neuronal para las lenguas románicas de la península Ibérica*, proyecto PID2021-124663OB-I00 y *Grandes modelos del lenguaje para la traducción de lenguas románicas de la península Ibérica con pocos recursos*, proyecto PID2024-158157OB-C33 financiados por MCIN /AEI /10.13039/501100011033 / FEDER, UE.

Bibliografía

Abdurakhmonova, N., A. Z. Mohirdev, M. Salokhiddinov, A. Narzullayev, y A. Gatiatullin. 2024. Nllb-based uzbek nmt: Leveraging multisource data. En *2024 9th International Conference on Computer Science and Engineering (UBMK)*, páginas 1–5. IEEE.

Alves, D. M., J. Pombal, N. M. Guerreiro, P. H. Martins, J. Alves, A. Farajian, B. Peters, R. Rei, P. Fernandes, S. Agrawal, P. Colombo, J. G. C. de Souza, y A. F. T. Martins. 2024. Tower: An

[6] `https://tan-ibe.github.io`

open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Artetxe, M., G. Labaka, y E. Agirre. 2019. An effective approach to unsupervised machine translation. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 194–203.

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

Bañón, M., P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, P. Koehn, R. Sennrich, y A. Waites. 2020. Paracrawl: Web-scale parallel corpora for the masses. En *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, páginas 3555–3564, Marseille, France. European Language Resources Association (ELRA).

Bié, L., A. Cerdà-i Cucó, H. Degroote, A. Estela, M. García-Martínez, M. Herranz, A. Kohan, M. Melero, T. O'Dowd, S. O'Gorman, y others. 2020. Neural translation for the european union (nteu) project. En *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, páginas 477–478.

Burchell, L., O. de Gibert, N. Arefyev, M. Aulamo, M. Bañón, P. Chen, M. Fedorova, L. Guillou, B. Haddow, J. Hajič, J. Helcl, E. Henriksson, M. Klimaszewski, V. Komulainen, A. Kutuzov, J. Kytöniemi, V. Laippala, P. Mæhlum, B. Malik, F. Mehryary, V. Mikhailov, N. Moghe, A. Myntti, D. O'Brien, S. Oepen, P. Pal, J. Piha, S. Pyysalo, G. Ramírez-Sánchez, D. Samuel, P. Stepachev, J. Tiedemann, D. Variš, T. Vojtěchová, y J. Zaragoza-Bernabeu. 2025. An expanded massive multilingual dataset for high-performance language technologies (hplt). En *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 17452–17485, Vienna, Austria, Julio. Association for Computational Linguistics.

Caswell, I., C. Chelba, y N. Constant. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. En *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, páginas 6588–6603, Barcelona, Spain (Online). International Committee on Computational Linguistics.

CE. 1992. European charter for regional or minority languages. European Treaty Series No. 148.

Costa-jussà, M. R., J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, y J. Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

de Dios-Flores, I., C. Magariños, A. I. Vladu, J. E. Ortega, J. R. Pichel, M. García, P. Gamallo, E. Fernández Rei, A. Bugarín, M. González González, S. Barro, y X. L. Regueira. 2022. The nós project: Opening routes for the galician language in the field of language technologies. En *Proceedings of the TDLE Workshop @ LREC 2022*, páginas 52–61, Marseille, France. European Language Resources Association (ELRA).

Edunov, S., M. Ott, M. Auli, y D. Grangier. 2018. Understanding back-translation at scale. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 489–500.

Fan, A., S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, y et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Firat, O., K. Cho, y Y. Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. En *Proceedings of NAACL-HLT*.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, y F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. En *Machine Translation Summit XIII: Proceedings of the Thirteenth Machine Translation Summit*, páginas 127–134, Xiamen, China. Asia-Pacific Association for Machine Translation (AAMT).

Garcia Gilabert, J., X. Liao, S. Da Dalt, E. Bohman, A. Mash, F. De Luca Fornaciari, I. Baucells, J. Llop, M. Claramunt, C. Escolano, y M. Melero. 2025. From SALAMANDRA to SALAMANDRATA: BSC submission for WMT25 general machine translation shared task. En B. Haddow T. Kocmi P. Koehn, y C. Monz, editores, *Proceedings of the Tenth Conference on Machine Translation*, páginas 614–637, Suzhou, China, Noviembre. Association for Computational Linguistics.

Gonzalez-Agirre, A., M. Marimon, C. Rodriguez-Penagos, J. Aula-Blasco, I. Baucells, C. Armentano-Oller, J. Palomar-Giner, B. Kulebi, y M. Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of catalan. En *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, páginas 2556–2566, Torino, Italia, Mayo. ELRA and ICCL.

Guzmán, F., P.-J. Chen, M. Ott, J. Pino, G. Lample, M. Lewis, V. Chaudhary, A. Fan, S. Bhosale, N. Goyal, A. El-Kishky, G. Wenzek, A. Chaudhary, P. Ng, J. Gu, S. Edunov, A. Baevski, M. Auli, y A. Mohamed. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 6098–6111. Association for Computational Linguistics.

Huck, M. y et al. 2020. Client-side neural machine translation for european languages. En *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, páginas 47–55.

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, y A. Birch. 2018a. Marian: Fast neural machine translation in C++. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, páginas 116–121, Melbourne, Australia. Association for Computational Linguistics.

Junczys-Dowmunt, M., K. Heafield, H. Hoang, R. Grundkiewicz, y A. Aue. 2018b. Marian: Cost-effective high-quality neural machine translation in c++. En *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, páginas 129–135. Association for Computational Linguistics.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koehn, P. y R. Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*. arXiv:1706.03872.

Kornai, A. 2013. Digital language death. *PLOS ONE*, 8(10):e77056.

Kudo, T. y J. Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Lample, G., M. Ott, A. Conneau, L. Denoyer, y M. Ranzato. 2018. Phrase-based & neural unsupervised machine translation. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 5039–5049.

Li, Z., X. Liu, D. F. Wong, L. S. Chao, y M. Zhang. 2022. Consisttl: Modeling consistency in transfer learning for low-resource neural machine translation. En *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 12543–12557.

Nguyen, T. Q. y D. Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation.

En *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, páginas 296–301.

Oliver, A. y S. Álvarez. 2023. Filtering and rescoring the CCMatrix corpus for neural machine translation training. En M. Nurminen J. Brenner M. Koponen S. Latomaa M. Mikhailov F. Schierl T. Ranasinghe E. Vanmassenhove S. A. Vidal N. Aranberri M. Nunziatini C. P. Escartín M. Forcada M. Popovic C. Scarton, y H. Moniz, editores, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, páginas 39–45, Tampere, Finland, Junio. European Association for Machine Translation.

Pan, X., M. Wang, L. Wu, y L. Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. En *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, páginas 7101–7117.

Papineni, K., S. Roukos, T. Ward, y W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, páginas 311–318.

Pérez-Ortiz, J. A., F. Sánchez-Martínez, V. M. Sánchez-Cartagena, M. Esplà-Gomis, A. Galiano-Jiménez, A. Oliver, C. Aventín-Boya, A. Pardos, C. Valdés, J. L. S. Socasau, y others. 2024. Expanding the flores+ multilingual benchmark with translations for aragonese, aranese, asturian, and valencian. En *Proceedings of the Ninth Conference on Machine Translation*, páginas 547–555.

Popović, M. 2016. chrf deconstructed: beta parameters and n-gram weights. En *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, páginas 499–504.

Post, M. 2018. A call for clarity in reporting BLEU scores. En *Proceedings of the Third Conference on Machine Translation: Research Papers*, páginas 186–191, Belgium, Brussels, Octubre. Association for Computational Linguistics.

Rei, R., C. Stewart, A. C. Farinha, y A. Lavie. 2020. Comet: A neural framework for mt evaluation. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 2685–2702.

Reimers, N. y I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sánchez-Martínez, F., J. A. Pérez-Ortiz, A. Galiano-Jiménez, y A. Oliver. 2024. Findings of the wmt 2024 shared task translation into low-resource languages of spain: Blending rule-based and neural systems. En *Proceedings of the Ninth Conference on Machine Translation*, páginas 684–698.

Sant, A., D. Bardanca, J. R. Pichel Campos, F. De Luca Fornaciari, C. Escolano, J. Garcia Gilabert, P. Gamallo, A. Mash, X. Liao, y M. Melero. 2024. Training and fine-tuning NMT models for low-resource languages using apertium-based synthetic corpora. En B. Haddow T. Kocmi P. Koehn, y C. Monz, editores, *Proceedings of the Ninth Conference on Machine Translation*, páginas 925–933, Miami, Florida, USA, Noviembre. Association for Computational Linguistics.

Schwenk, H., G. Wenzek, S. Edunov, É. Grave, A. Joulin, y A. Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. En *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 6490–6500.

Sellam, T., D. Das, y A. Parikh. 2020. Bleurt: Learning robust metrics for text generation. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 7881–7892.

Sennrich, R., B. Haddow, y A. Birch. 2016. Improving neural machine translation mo-

dels with monolingual data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, páginas 86–96.

Sharami, J. P. R., D. Sterionov, y P. Spronck. 2021. Selecting parallel in-domain sentences for neural machine translation using monolingual texts. *Computational Linguistics in the Netherlands Journal*, 11:213–230.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, y J. Makhoul. 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: Technical Papers*, páginas 223–231.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. En N. Calzolari K. Choukri T. Declerck M. U. Doğan B. Maegaard J. Mariani A. Moreno J. Odijk, y S. Piperidis, editores, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, páginas 2214–2218, Istanbul, Turkey, Mayo. European Language Resources Association (ELRA).

Tiedemann, J. y S. Thottingal. 2020. OPUS-MT – building open translation services for the world. En A. Martins H. Moniz S. Fumega B. Martins F. Batista L. Coheur C. Parra I. Trancoso M. Turchi A. Bisazza J. Moorkens A. Guerberof M. Nurminen L. Marg, y M. L. Forcada, editores, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, páginas 479–480, Lisboa, Portugal, Noviembre. European Association for Machine Translation.

UNESCO. 2003. Language vitality and endangerment. Informe técnico, UNESCO. Report submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages, Paris, 10–12 March 2003.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, y I. Polosukhin. 2017. Attention is all you need. En *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, páginas 6000–6010, Long Beach, CA, USA.

Wenzek, G., M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, y E. Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. En *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, páginas 4003–4012, Marseille, France. European Language Resources Association (ELRA).

Xu, C., Y. Cui, J. Liu, y et al. 2024. Scaling neural machine translation to 200 languages. *Nature*.

Zoph, B., D. Yuret, J. May, y K. Knight. 2016. Transfer learning for low-resource neural machine translation. En J. Su K. Duh, y X. Carreras, editores, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, páginas 1568–1575, Austin, Texas, Noviembre. Association for Computational Linguistics.