

Capítulo 5

Revisión sistemática de literatura: Traductores automáticos basados en inteligencia artificial para lenguas indígenas con escasos recursos lingüísticos, utilizando el Modelo de Red Neuronal Transformer

*Rolando Bautista Morales¹
Yobani Martínez Ramírez²
Alan Ramírez Noriega³
José Emilio Sánchez García⁴*

<https://doi.org/10.61728/AE24050050>

¹ rolandobautista@uaim.edu.mx, Universidad Autónoma Indígena de México

² yobani@uas.edu.mx, Universidad Autónoma de Sinaloa

³ alandramireznoriega@uas.edu.mx, Universidad Autónoma de Sinaloa

⁴ esanchez@uaim.edu.mx, Universidad Autónoma Indígena de México

Resumen

El presente artículo tiene como propósito realizar una revisión sistemática de literatura (RSL) con la finalidad de explorar las nuevas tecnologías que están emergiendo en el campo de la inteligencia artificial (IA) (Cabe mencionar que estas) están enfocadas en redimir a aquellas lenguas que presentan escasos recursos lingüísticos (ERL) y que, por lo anterior, se encuentran en peligro de extinción. Conviene destacar que la traducción automática (TA) y las redes neuronales son los principales subcampos de la IA que han aportado posibles soluciones a este problema. Este trabajo se enfoca en analizar a aquellas investigaciones que utilizan el modelo de red neuronal llamada Transformer. Para ello, se implementa el método estándar de RSL, el cual examina una serie de 60 documentos científicos que han sido publicados en los últimos años. A grandes rasgos, los objetivos de este estudio son identificar los procesos de construcción de los traductores automáticos neuronales (TAN), precisar las herramientas que se utilizan en cada una de sus etapas, y, por último, evaluar el grado de eficacia alcanzada. Los resultados de este trabajo facilitaron una visión general del estado actual de la investigación en este campo.

Introducción

En la actualidad, la traducción automática es una herramienta importante para facilitar la comunicación entre personas que hablan diferentes idiomas, sin embargo, la eficacia de los traductores automáticos puede ser limitada por la falta de disponibilidad de recursos lingüísticos, este es el caso de las lenguas minoritarias. Los autores (Chakravarthi et al., 2021) plantean que, “debido a la falta de recursos digitales, muchas de estas lenguas podrían extinguirse, y por ende se perdería la conexión con la cultura de los pueblos y las características de las lenguas”.

Cabe mencionar que Rai et al., (2023) exponen en su investigación que recientemente las aplicaciones de aprendizaje automático y las técnicas de aprendizaje profundo han llamado mucho la atención en prácticamente todos los dominios de la investigación, y la ingeniería de

software no es la excepción. Así mismo, Marcus (2018) dice que los algoritmos de aprendizaje, como aprendizaje profundo, requieren datos masivos para su entrenamiento. Ante estos retos que se anteponen a la traducción automática para lenguas de escasos recursos en Vaswani et al., (2017) se plantea un nuevo modelo de red neuronal que cambia literalmente la concepción que se tiene de lo que era capaz de lograr la inteligencia artificial (IA) en el campo del procesamiento de lenguaje natural (PLN) y de los TA. Este modelo de red neuronal tiene como principal objetivo la sustitución de las redes neuronales recurrentes como las Long Short-Term Memory (LSTMs) que se venían usando en el campo del PLN, por las denominadas Transformer. Este modelo plantea ser una solución prometedora para mejorar la calidad y precisión de los TA con escasos recursos.

En Alegria et al. (2011) los autores proponen una tipología lingüística de seis niveles para desarrollar tecnologías lingüísticas que pudieran utilizarse en varios cientos de lenguas. El primer nivel corresponde a las lenguas con más recursos; el segundo, a las que figuran entre las 10 lenguas más utilizadas en la red. El tercer nivel son las lenguas que cuentan con algún tipo de recurso en PLN. El cuarto nivel incluye las lenguas que tienen algún tipo de recurso léxico. Las lenguas que tienen un sistema de escritura, pero no en formato digital están en el quinto nivel. El último nivel es significativo, ya que incluye las lenguas orales que no tienen un sistema de escritura propio.

Las lenguas que carecen de corpus paralelos extensos se conocen como lenguas con pocos recursos o con recursos insuficientes (Jimerison y Prud'Hommeaux, 2018). Para este trabajo, definimos como lenguas con pocos recursos las que se encuentran en los niveles tercero, cuarto, quinto y sexto, porque los retos a los que se enfrentan son de un nivel técnico y no de naturaleza social.

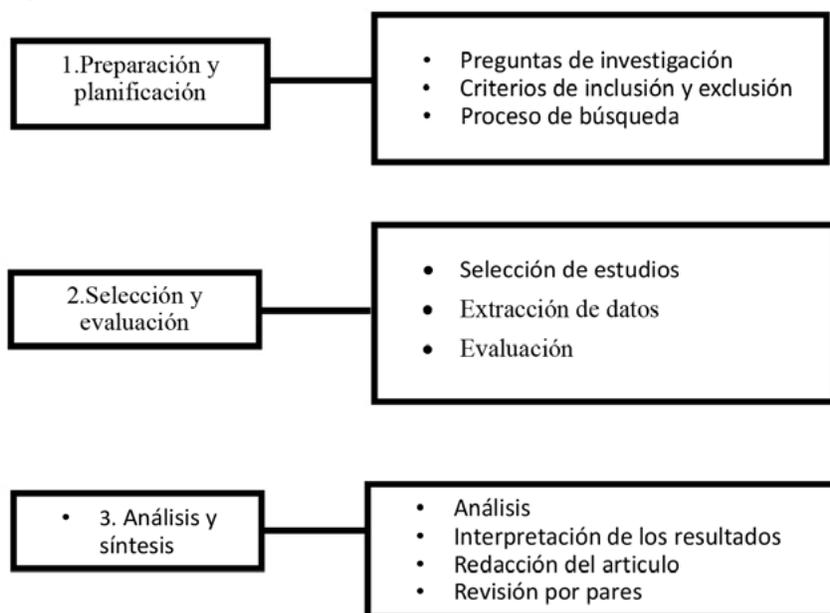
En este trabajo de investigación se presenta una revisión sistemática de literatura (RSL) que se centra en las investigaciones que se ha venido publicando en relación con la TAN que utiliza el modelo Transformer con escasos recursos lingüísticos (ERL), sean nacionales o internacionales. Se espera que los resultados obtenidos de esta revisión sirvan como una guía útil para las futuras investigaciones y desarrollos en el área de la TAN con ERL para lenguas minoritarias.

Método

La metodología utilizada en la presente investigación para la RSL, está basada en la propuesta de Kitchenhan (Budgen y Brereton, 2006), la cual fue aplicada en las áreas de investigación de software. Este método implica una búsqueda exhaustiva de artículos científicos y demás documentos relevantes en la materia.

Estos autores definen a la RSL como un medio para evaluar e interpretar todas las investigaciones disponibles relevantes para una determinada pregunta de investigación, área temática o fenómeno de interés. Esta tiene como objetivo presentar una evaluación justa de un tema de investigación, utilizando una metodología fiable, rigurosa y comprobable (Lerandi et al., 2020).

Figura 1: *Proceso de revisión sistemática de literatura.*



Las preguntas de investigación

Las preguntas de investigación que se abordan en este estudio son:

RQ1. ¿Qué tipo de tecnologías utilizan los traductores automáticos para lenguas con escasos recursos lingüísticos?

RQ2. ¿Cuál es el proceso de construcción de un traductor automático con el modelo Transformer?

RQ3. ¿Cuál es el grado de eficacia alcanzada en las traducciones?

Criterios de inclusión y exclusión

Para asegurar la rigurosidad de la RSL, se establecieron criterios de inclusión y de exclusión para la selección de los artículos. En la primera fase, se incluyeron todos los artículos que mencionaran los términos: traductor, redes neuronales, aprendizaje automático, lenguas indígenas o minoritarias, independientemente del idioma en el que estuvieran escritos. En la segunda fase, los criterios de inclusión y exclusión como se muestra en la Tabla 1, se eligieron aquellos artículos que trabajaban específicamente con IA.

Tabla 1: *Criterios de inclusión y exclusión*

Inclusión	Exclusión
Publicados de enero del 2015 hasta diciembre del 2022	Publicados antes del 2015
Traductores automáticos con IA	Traductores sin IA
Traductores automáticos para lenguas con escasos recursos lingüísticos	Traductores automáticos con muchos recursos lingüísticos
Artículos o tesis en inglés y español	Artículos que no estén en inglés o español
Modelos que trabajan con la arquitectura Transformer	Modelos que no trabajan con la arquitectura Transformer

Proceso de búsqueda

En la primera fase, se realizó una búsqueda exhaustiva en diferentes bases de datos de documentos incluyendo artículos científicos y tesis relacionadas con la aplicación de la IA en traductores automáticos para lenguas minoritarias o de escasos recursos. Se utilizaron cinco bases de datos diferentes para la búsqueda: Aclantogy, Google Académico, Google, The Nacional Library of Medicine, ResearchGate y Hindawi. Se encontraron un total de sesenta documentos relevantes.

Para ser más selectivos y darle un enfoque hacia el estudio de aplicaciones con IA en traductores automáticos, se evaluó cada artículo en función de su título y resumen. Es por ello por lo que se seleccionaron los que parecían más relevantes para el estudio. De los sesenta artículos iniciales, solo treinta cumplieron con los criterios de inclusión, excluyendo aquellos que no utilizaban ningún tipo de IA en sus traductores.

Posteriormente, en la tercera fase de la revisión, se evaluaron los treinta artículos de una forma más exhaustiva, profunda y analítica, de lo cuales se seleccionó solamente quince documentos, excluyendo aquellos que cumplen con los demás requisitos de inclusión ya mencionadas anteriormente. En este sentido, se seleccionaron aquellos que daban información sobre el desarrollo de traductores con IA y redes neuronales para lenguas de escasos recursos lingüísticos.

De los quince artículos analizados, se excluyeron cuatro debido a su enfoque en el reconocimiento automático de voz (RAV). De estos, únicamente once cumplieron con los criterios previamente mencionados de inclusión y exclusión. Los detalles de los artículos seleccionados se presentan en la Tabla 2.

Tabla 1: Criterios de inclusión y exclusión

No	Tipo	Título	Base de datos	Cita
1	Artículo	Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut DOI: 10.18653/v1/2020.coling-main.410	Aclan- tology	(Le y Sadat, 2020)
2	Artículo	Ayuuk – Spanish Neural Machine Translator DOI: 10.18653/v1/2021.americasnlp-1.19	Aclan- tology	(Zacarías y Meza, 2021)
3	Artículo	NRC-CNRC Machine Translation Systems for the 2021 AmericasNLP Shared Task DOI: 10.18653/v1/2021.americasnlp-1.25	Aclan- tology	(Knowles et al., 2021)
4	Artículo	The REPU CS' Spanish–Quechua Submission to the Americas NLP 2021 Shared Task on Open Machine Translation 10.18653/v1/2021.americasnlp-1.27	Aclan- tology	(Veliz, 2021)
5	Artículo	IndT5: un transformador de texto a texto para 10 lenguas indígenas DOI: 10.18653/v1/2021.americasnlp-1.30	Aclan- tology	(Moatez et al., 2021)
6	Artículo	Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri DOI: 10.18653/v1/2020.coling-main.351	Google	(Feldman y Co- to-Sola- no, 2020)
7	Tesis	Traducción automática neuronal para lengua nativa peruana	Google Acadé- mico	(Huarca- ya, 2020)
8	Artículo	Improving Neural Machine Translation of Indigenous Languages with Multilingual Transfer Learning DOI: https://doi.org/10.48550/arXiv.2205.06993	Google Acadé- mico	(Chen y Ab- dul-Ma- geed, 2022)
9	Artículo	Machine Translation System Using Deep Learning for English to Urdu DOI: https://doi.org/10.1155/2022/7873012	National Library of Me- dicine	(Andrabi y Wahid, 2022)

No	Tipo	Título	Base de datos	Cita
10	Artículo	English-Chinese Machine Translation Based on Transfer Learning and Chinese-English Corpus DOI: https://doi.org/10.1155/2022/1563731	Hindawi	(Xu, 2022)
11	Artículo	Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages DOI: 10.18653/V1/2020.FINDINGS-EM-NLP.195	Aclan- tology	(Nekoto et al., 2020)

Finalmente, quedaron diez artículos científicos y una tesis. Estas investigaciones están orientadas específicamente en el desarrollo de la IA en traductores automáticos para lenguas minoritarias o de escasos recursos que trabajan con el modelo neuronal Transformer. Estos documentos fueron evaluados a detalle para extraer información importante sobre la relevancia y eficacia de los TA.

Resultados

En esta sección se indican las herramientas que utilizaron investigadores en el proceso de construcción de los traductores automáticos.

Los resultados revelan un patrón común que fue aplicado en el desarrollo de los traductores automáticos. En la mayoría de los casos, se observó que los traductores automáticos fueron construidos siguiendo un proceso que consta de tres fases clave: la alineación automática, la tokenización y el entrenamiento de la red neuronal.

En la primera, la alineación automática es crucial para establecer correspondencias entre las frases y palabras de los idiomas de origen y destino. En esta fase se encontró que varios estudios como los de Le y Sadat (2020), Moatez et al. (2021) y Feldman y Coto-Solano (2020), utilizaron la herramienta WordPiece como parte de la fase de alineación automática en el desarrollo de los traductores automáticos.

Por otro lado, en la primera fase los autores Zacarías y Meza (2021) emplean una herramienta llamada YASA (Yet Another Fast, Robust and Open Source Sentence Aligner) para realizar la alineación automática en su traductor automático. YASA es una herramienta de código abierto que se ha desarrollado específicamente para lograr una alineación de frases rápidas y confiables en el contexto de la traducción automática (Lamraoui y Langlais, 2013).

Estos instrumentos demuestran la diversidad de métodos y herramientas utilizados en la fase de alineación automática. Este hallazgo resalta la importancia de evaluar y comparar diferentes enfoques para lograr una alineación precisa y efectiva en el desarrollo de traductores automáticos de lenguas minoritarias.

En la segunda fase, la tokenización, se refiere a la segmentación de los textos en unidades más pequeñas, como palabras o subpalabras que se utilizan como unidades de entrada y salida para el traductor automático.

En esta fase, Le y Sadat, (2020), Zacarías y Meza (2021) y Feldman y Coto-Solano (2020b) usaron la herramienta Subword-nmt library, esta herramienta es ampliamente utilizada en la tokenización de textos y está basada en el enfoque de subpalabras. Esta librería contiene scripts de preprocesamiento para segmentar texto en unidades de subpalabras. Su objetivo principal es facilitar la reproducción de experimentos sobre la traducción automática neuronal con unidades de subpalabras (Comunity Python, 2023).

Los autores Moatez et al. (2021) utilizaron la herramienta WordPiece, que utiliza un algoritmo de tokenización que Google desarrolló para BERT (Comunity Hugging Face, 2022).

Por otro lado, el autor Veliz (2021) utilizó la herramienta Byte-Pair Encoding (BPE). Esta utiliza un método de tokenización que divide palabras en subunidades basadas en secuencias de bytes. BPE se desarrolló inicialmente como un algoritmo para comprimir textos, y luego fue utilizada por la empresa OpenAI para la tokenización al entrenar diferentes modelos. También es utilizado por muchos modelos de Transformer, incluidos GPT (Generative Pre-trained Transformer), GPT-2, RoBERTa (Robustly Optimized BERT Pretraining Approach), BART

(Bidirectional and Auto-Regressive Transformers) y DeBERTa (Decoding-enhanced BERT with Disentangled Attention) (Comunity Hugging Face, 2022).

El autor (Huarcaya, 2020) emplea la herramienta Word embeddings. Esta tiene una técnica que utiliza representaciones vectoriales para palabras individuales, lo que ayuda a capturar la semántica y similitud entresignos lingüísticos. Las incrustaciones de palabras nos brindan una manera de usar una representación densa y eficiente en la que palabras similares tienen una codificación similar (TensorFlow, 2022).

La utilización de estas herramientas de tokenización resaltan la importancia del uso de enfoques de segmentación adecuada en el procesamiento de textos en los traductores automáticos, en donde es importante aclarar que cada una de estas contribuyen a mejorar la calidad de las traducciones automáticas, especialmente en el contexto de lenguas minoritarias o de escasos recursos.

Finalmente, la tercera fase implica el entrenamiento de la red neuronal. En esta etapa, se utiliza un conjunto de datos de entrenamiento para ajustar los parámetros de la red neuronal y mejorar su capacidad de traducción. El entrenamiento se basa en algoritmos de aprendizaje automático que permiten a la red neuronal aprender patrones y mejorar su rendimiento con el tiempo. Estas herramientas desempeñaron un papel fundamental en el entrenamiento de las redes neuronales utilizadas en los sistemas de traducción automática.

Los autores Le y Sadat (2020), Veliz (2021) y Chen y Abdul-Mageed (2022) implementaron Marian NMT (Traducción automática neuronal). Esta es una herramienta de traducción automática neuronal desarrollada por el equipo de investigación de Marian. Esta tecnología utiliza una infraestructura flexible para el entrenamiento de la red neuronal. Los autores (Junczys-Dowmunt et al., 2018) plantean que Marian es un marco de Traducción Automática Neuronal eficiente y autónomo con un motor de diferenciación automático integrado basado en grafos gráficos de cálculo dinámicos.

Por otra parte, los autores Zacarías y Meza (2021) y Huarcaya (2020) emplearon JoeyNMT, que es una librería de código abierto para la traducción automática neuronal desarrollada en Python. JoeyNMT

proporciona una implementación modular y flexible de los sistemas de traducción automática neuronal, lo que facilita el entrenamiento y la evaluación de modelos de traducción automática. Los teóricos (Kreutzer et al., 2019) comentan que a pesar de su enfoque en la simplicidad, JoeyNMT admite arquitecturas clásicas (RNN, transformadores), búsqueda rápida, vinculación de peso y más, y logra un rendimiento comparable contra otras herramientas más complejas.

Los investigadores Andrabi y Wahid (2022) hacen uso de KERAS. Esta es una librería de aprendizaje profundo muy popular y de código abierto que se utiliza ampliamente en la comunidad de la inteligencia artificial. Keras: Deep Learning Para Humanos, (2023) dice que el propósito de Keras es dar una unfair advantage a cualquier desarrollador que busque enviar aplicaciones impulsadas por Machine Learning. Keras se centra en la velocidad de depuración, la elegancia y la concisión del código, la capacidad de mantenimiento y la capacidad de implementación.

En esta fase Feldman y Coto-Solano (2020) implementa la técnica de retrotraducción iterativa para la generación de oraciones sintéticas.

Discusiones

En esta sección se indica lo que recomiendan los autores para futuras investigaciones.

La generación de oraciones sintéticas a través de la retrotraducción iterativa es de suma importancia, porque proporciona una mayor variedad de ejemplos y ayuda a entrenar los modelos de traducción automática en una amplia gama de contextos lingüísticos. En Duy et al. (2018) el autor dice que este es un método para generar datos paralelos sintéticos cada vez mejores a partir de datos monolingües para entrenar sistemas neuronales de traducción automática. Además, en (Chakravarthi et al., 2021) el autor plantea que la traducción automática neuronal reciente es completamente integral, pero adolece de escasez de datos cuando se trata de lenguas ricas morfológicamente o con pocos recursos. Sin embargo, dice que se ha demostrado que el cambio de código basado en la ortografía o en la traducción automática neuronal basada

en caracteres mejora los resultados de forma significativa. Al considerar la implementación de traducción automática basada en caracteres, se puede mejorar significativamente la calidad y precisión de las traducciones, especialmente en lenguas con escasos recursos donde la disponibilidad de recursos léxicos puede ser limitada.

En Feldman y Coto-Solano (2020) los autores detectaron en un experimento que realizaron que parte del corpus que utilizaron fueron libros de texto con reglas gramaticales y considera que estos son ricos en ejemplos que los modelos aprovecharon para mejorar la traducción.

La implementación de alguna de estas herramientas de entrenamiento destaca la importancia de contar con herramientas y librerías especializadas en el desarrollo de sistemas de traducción automática. En la Tabla 3 se analizaron las herramientas que se implementaron en cada fase de la construcción de los traductores automáticos.

Tabla 3 : *Herramientas utilizadas en las 3 fases en la RSL*

Artículos	Fase 1. Alineación automática	Fase 2 Tokeniza- ción	Fase 3. Entrena- miento	Eficacia métrica	Número de oraciones
(Le y Sadat, 2020)	LSTM (Long Short-Term Memory)	Su- bword-nmt library4	Marian NMT	Bajo	No Indica
(Zacaría y Meza, 2021)	YASA (Yet Another Fast, Robust and Open Source Sentence Aligner)	Su- bword-nmt library4	JoeyNMT5	Bajo	6000
(Knowles et al., 2021)	No Específica	No Específica	No Específica	No Específica	No Indica
(Moatez et al., 2021)	Segmentación de palabras basado en BERT	Wordpiece	indT5	Baja	No Específica

Artículos	Fase 1. Alineación automática	Fase 2 Tokeniza- ción	Fase 3. Entrena- miento	Eficacia métrica	Número de oraciones
(Veliz, 2021)	Byte-Pair Encoding (BPE)	BPE (Byte Pair Encod- ing)	Marian NMT	Media	32000
(Feldman y Coto-Sola- no, 2020)		Modelo entrenado desde 0, en PyTorch (Uso de oraciones sintéticas)		Media/ Alta	5923
(Huarcaya, 2020)	Word embedding	BPE (Byte Pair Encod- ing)	JoeyNMT5	Media	119000
(Chen y Abdul-Ma- geed, 2022)	No Especí- fica	No Especí- fica	Marian NMT	Baja	No Especí- fica
(Andrabi y Wahid, 2022)	Word embedding	No Especí- fica	KERAS	Media/ Alta	30923
(Xu, 2022)	No Especí- fica	No Especí- fica	No Especí- fica	Alta	Más de 100 mil
(Nekoto et al., 2020)	No Especí- fica	No Especí- fica	No Especí- fica	No Especí- fica	Más de 100 mil
Propuesta (Rolando, 2023)	Word embedding	BPE (Byte Pair Encod- ing)	KERAS (Uso de oraciones sintéticas)	Alta	6000 (Incluir diferentes contextos de las ora- ciones)

Las arquitecturas como las LSTM y BERT, cuentan con librerías propias para la fase de alineación esto significa que no dependen de herramientas específicas para realizar este proceso; de igual forma JoeyNMT5 y Marian NMT, son arquitecturas que cuentan con diferentes librerías y herramientas para llevar a cabo el proceso de construcción de un traductor automático.

Conclusiones

En conclusión, los estudios seleccionados presentan una clara tendencia hacia el uso de tecnologías como PLN, aprendizaje automático y redes neuronales, estas tecnologías demostraron ser efectivas en la traducción automática, especialmente en el contexto de lenguas minoritarias o de escasos recursos lingüísticos.

Sin embargo, para determinar cuál de las herramientas mencionadas anteriormente son las más óptimas para implementarlas en futuros desarrollos de traductores automáticos con escasos recursos depende directamente del enfoque y de los recursos lingüísticos con los que dispone cada investigador. Cabe resaltar que estas herramientas han demostrado ser muy efectivas, pero es fundamental considerar las características específicas de cada proyecto y tener en cuenta las limitaciones con las que se llegara a tener.

Para las lenguas que se ubican en el quinto y sexto nivel, como es el caso del proyecto de Zacarías y Meza (2021) que entreno a su red neuronal con seis mil pares de oraciones, además junto con los autores Feldman y Coto-Solano (2020) y Andrabi y Wahid (2022) concluyeron que es necesario trabajar con corpus de datos más extensos para mejorar el rendimiento de los traductores automáticos.

Es por ello por lo que el desarrollo de software con IA para la traducción de lenguas minoritarias o de escasos recursos lingüísticos es de suma importancia debido a los diversos beneficios que aporta a nivel cultural, social y académico.

Estas nuevas tecnologías son de gran importancia, dado que contribuye a preservar la diversidad cultural, facilitar el acceso a la información, promover la conectividad global y apoyar la investigación lingüística, lo cual tiene impactos positivos en las comunidades indígenas y en la sociedad en general.

Referencias

- Alegria, I., Artola, X., De Ilarraza A. D., & Sarasola, K. (2011). *Strategies to develop language technologies for less-resourced languages based on the case of Basque*.
- Andrabi, S. A. B., & Wahid, A. (2022). Machine Translation System Using Deep Learning for English to Urdu. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/7873012>
- Budgen, D., & Brereton, P. (2006). Performing systematic literature reviews in software engineering. *Proceedings of the 28th International Conference on Software Engineering*, 1051–1052. <https://doi.org/10.1145/1134285.1134500>
- Chakravarthi, B. R., Rani, P., Arcan, M., & McCrae, J. P. (2021). A Survey of Orthographic Information in Machine Translation. *SN Computer Science*, 2(4), 330. <https://doi.org/10.1007/s42979-021-00723-4>
- Chen, W.-R., & Abdul-Mageed, M. (2022). *Improving Neural Machine Translation of Indigenous Languages with Multilingual Transfer Learning*. <http://arxiv.org/abs/2205.06993>
- Comunity Hugging Face. (2022). <https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt>
- Comunity python. (2023). <https://pypi.org/project/subword-nmt/>
- Duy, C., Hoang, V., Koehn, P., Haffari, G., & Cohn, T. (2018). Iterative Back-Translation for *Neural Machine Translation*. 18–24.
- Feldman, I., & Coto-Solano, R. (2020a). *Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri*. Online.
- Feldman, I., & Coto-Solano, R. (2020b). Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri. *Proceedings of the 28th International Conference on Computational Linguistics*, 3965–3976. <https://doi.org/10.18653/v1/2020.coling-main.351>
- Huarcaya, D. (2020). *Traducción automática neuronal para lengua nativa peruana*.

- Ierandi, C., Orihuela Espina, L., Jurado Flores, I., Rodríguez del Nozal, Á., & Tapia Córdoba, A. (2020). *Revisión sistemática de la literatura en ingeniería de sistemas. Caso práctico: técnicas de estimación distribuida de sistemas ciberfísicos*. 84–91. <https://doi.org/10.17979/spudc.9788497497749.0084>
- Incrustaciones de palabras, TensorFlow*. (2022). https://www.tensorflow.org/text/guide/word_embeddings?hl=es-419
- Jimerson, R., & Prud'Hommeaux, E. (2018). *ASR for Documenting Acutely Under-Resourced Indigenous Languages*. <https://aclanthology.org/L18-1657>
- Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., & Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. *Proceedings of ACL 2018, System Demonstrations*, 116–121. <https://doi.org/10.18653/v1/P18-4020>
- Keras: Deep Learning para humanos. (2023). <https://keras.io/>
- Knowles, R., Stewart, D., Larkin, S., & Littell, P. (2021). NRC-CNRC Machine Translation Systems for the 2021 AmericasNLP Shared Task. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 224–233. <https://doi.org/10.18653/v1/2021.americasnlp-1.25>
- Kreutzer, J., Bastings, J., & Riezler, S. (2019). Joey NMT: A Minimalist NMT Toolkit for Novices. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 109–114. <https://doi.org/10.18653/v1/D19-3019>
- Lamraoui, F., & Langlais, P. (2013). *Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment?* <http://www.statmt.org/>
- Le, N. T., & Sadat, F. (2020). *Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut*. Online. <https://github.com/huggingface/transformers>
- Marcus, G. (2018). *Deep Learning: A Critical Appraisal*.
- Moatez, E., Nagoudi, B., Chen, W.-R., Abdul-Mageed, M., & Cavuso-

- glu, H. (2021). *IndT5: A Text-to-Text Transformer for 10 Indigenous Languages*. <https://github.com/UBC-NLP/IndT5>
- Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohunge, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., Freshia, S., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Meressa, M., Adeyemi, M., Mokgesi-Seling, M., Okegbemi, L., ... Bashir, A. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, 9, 2144–2160. <https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.195>
- Rai, S., Belwal, R. C., & Gupta, A. (2023). Is the Corpus Ready for Machine Translation? A Case Study with Python to Pseudo-Code Corpus. *Arabian Journal for Science and Engineering*, 48(2), 1845–1858. <https://doi.org/10.1007/s13369-022-07049-0>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, P., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems, 2017-December*, 5999–6009. <https://arxiv.org/abs/1706.03762v5>
- Veliz, O. M. (2021). The REPU CS' Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation. *Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas, AmericasNLP 2021*, 241–247. <https://doi.org/10.18653/V1/2021.AMERICASNLP-1.27>
- Xu, B. (2022). English-Chinese Machine Translation Based on Transfer Learning and Chinese-English Corpus. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/1563731>
- Zacarías, D., & Meza, I. (2021). *Ayuuk-Spanish Neural Machine Translator*. <https://github.com/anoidgit/yasa>

