DOI: https://doi.org/10.18485/beoiber.2023.7.1.13

Ernesto Llerena García[1]
Universidad de Córdoba
Colombia

SOFTWARE TRADUCTOR ESTADÍSTICO DE LENGUAS NATIVAS DE COLOMBIA: CASO LENGUA EMBERA KATÍO

Resumen

Este proyecto parte de la necesidad de comprender y reconocer las culturas étnicas, más específicamente la cultura embera katío que habita en la parte alta del río Sinú, Córdoba, Colombia. Actualmente se han visto obligados a desplazarse de ese territorio por circunstancias sociales (guerrillas, paramilitares y el impacto ambiental de la represa de Urrá). Con base en lo que sustenta la Ley de Protección de las Lenguas Originarias, las culturas étnicas tienen derecho a la educación, la inclusión y en este proceso es necesario utilizar las nuevas Tecnologías de la Información y la Comunicación para difundir y promover la preservación de las lenguas. Debido al contacto entre lenguas y culturas, se propone la construcción de un software traductor de oraciones simples y complejas del español al idioma embera katío para facilitar la comunicación y preservación de este idioma. Se describe la Metodología de Investigación en Ciencias del Diseño (DSRM) para el desarrollo del traductor. Como resultado de la investigación se está desarrollando una versión beta del software traductor en la versión estadística. Se espera que con este tipo de herramienta se logre apoyar los procesos de revitalización de la lengua embera katío.

Palabras clave: software traductor automático, pueblos indígenas, inclusión educativa, embera katío.

STATISTICAL TRANSLATOR SOFTWARE FOR NATIVE LANGUAGES OF COLOMBIA: EMBERA KATIO LANGUAGE CASE

Abstract

This project stems from the need to understand and recognize ethnic cultures, more specifically the Embera Katío culture that inhabits the upper part of the Sinú river, Córdoba, Colombia. Currently, they have been forced to leave this territory due to social circumstances (guerrillas, paramilitaries and the environmental impact of the Urrá dam). Based on what sustains the Law of Protection of Original Languages, the ethnic cultures have a right to education, inclusion and in this process, it is necessary to use the new technologies of information and communication to disseminate and promote the preservation of languages. Due to the contact between languages and cultures, it is proposed the construction of a software

[1] lle55re19@yahoo.com

to translate simple and complex sentences from Spanish to the Embera Katío language to facilitate the communication and preservation of this language. The Research Methodology in Design Sciences (DSRM) for translator development is described. As a result of the investigation, a beta version of the translator software is being developed in the statistical version. This type of tool is expected to support the revitalization processes of the Embera Katío language.

Key words: automatic translator software, indigenous peoples, educational inclusion, Embera Katío.

1. Introducción

La informática ha hecho avances bastante significativos en el procesamiento y almacenamiento de información en los últimos años. Se han creado múltiples programas para cumplir diversas funciones como procesadores de texto, procesadores gráficos, programas multimedia, traductores automáticos, entre otros. La traducción automática es una de las nuevas tendencias para comprender nuevos idiomas, como el español, el inglés, entre otros y utilizar la Inteligencia Artificial en el entorno académico (Barret 2019: 2; Lu 2018: 1). Un ejemplo de traductores automáticos es el *Traductor de Google* (Google 2006) el cual es un gran sistema de traducción de muchas de las lenguas del mundo; sin embargo, hay muchas lenguas que no están incluidas dentro de este sistema. La gran mayoría de los traductores sólo se enfocan en traducir idiomas reconocidos y se han dejado atrás aquellos de las culturas étnicas. Es el caso de la lengua embera katío, la cual no cuenta con un software traductor alojado en Internet.

Existen diferencias culturales entre los pueblos indígenas embera, dependiendo del entorno en el que viven. Los embera se componen así de dos grupos principales, los dobidas y los eyabidas. El pueblo dobida vive alrededor del río; sus casas y huertas están a orillas del río y siempre están pescando. Los eyabida incluyen los embera katío y los embera chamí. El nombre catíos (katíos) es el nombre que se le da en las crónicas a un pueblo indígena de origen y lengua caribeña que habitaba en una amplia zona del centro oeste, actual provincia de Antioquia, Colombia, en la época de la Conquista (Simón 1953: 420). Según esta descripción, el pueblo katío tenía una lengua chocoana y una lengua caribeña. Los chamíes son pueblos de montaña, a diferencia de los embera sapiadaras, que viven en las selvas de las llanuras del Pacífico, y los embera katíos, que viven en las cuencas del Atrato y el Alto Sinú (Werner Cantor 2000: 127). Lingüísticamente, los emberas se clasifican en emberas de occidente y emberas de oriente (Llerena 1994: 437).

Según la UNESCO (2022), hoy en día la cultura embera katío está desapareciendo paulatinamente dentro de la sociedad colombiana. Es una población que, debido a las circunstancias sociales, se ha visto obligada a desplazarse de su entorno habitual. La población estimada de indígenas embera en Colombia es de más de 80.000 habitantes, ya

que también se encuentra población embera en Panamá y Ecuador. En el departamento de Córdoba existen más de 8.000 personas de esta comunidad. En general, la cultura embera ha tenido un contacto permanente entre lenguas y al mismo tiempo contacto entre las diversas culturas que componen el territorio colombiano.

El gobierno colombiano ha creado la *Ley para la Protección de las Lenguas Originarias* (Ministerio de Cultura 2010), en la cual se reconocen todos los derechos y beneficios que tienen las 65 lenguas nativas y 2 lenguas criollas de Colombia. En esta se refieren a la preservación de las culturas nativas con la ayuda de las nuevas tecnologías de la información y la comunicación, como la televisión, Internet, entre otras. En cumplimiento de la Ley, la Universidad de Córdoba (Colombia) ha realizado un convenio con la comunidad embera katío del Alto Sinú para que los estudiantes indígenas que acceden a la educación primaria, secundaria y superior utilicen un software como apoyo a la conservación y revitalización de su lengua y su cultura. Así, el objetivo principal de este proyecto de investigación es crear un software que permita traducir oraciones simples y complejas del español a la lengua embera katío y viceversa. Se espera que este software traductor español-embera katío sirva, no solo para aquellos estudiantes indígenas que aprenden español, sino también para todos aquellos que, de una forma u otra, tienen algún tipo de contacto con esta cultura.

Este trabajo de investigación se basa básicamente en las TIC en la educación, y en especial en las poblaciones indígenas (Becerra Cortés 2012: 4). Son varias las investigaciones que han trabajado el tema del uso de las TIC en la educación indígena; al revisar la literatura en relación con la creación de aplicaciones tecnológicas para este fin, encontramos que en Colombia se han realizado varios proyectos a nivel nacional utilizando las TIC (Rojas et al. 2018: 881) pero no existen traductores de lenguas indígenas que utilicen información lingüística (RBMT), estadística (SMT) o basada en redes neuronales (NMT). A nivel internacional, en Centroamérica existe el *Software Traductor Español a Náhuatl* (Hernández 2015), el cual busca preservar el idioma náhuatl y ofrecer un medio que permita a los niños tener acceso a grandes cantidades de libros traducidos a su idioma de manera automática. En Suramérica encontramos el *Traductor Español-Guaraní, Guaraní-Español* (TraductorPro.com), el cual permite traducir palabras en ambos idiomas. En general, la creación de traductores se ha enfocado en lenguas SAE (Standard Average European) y en lenguas de África, América, Asia y Oceanía, como se puede evidenciar en el caso del traductor de Google, el cual soporta 133 lenguas. En la actualidad, existen proyectos con bastante desarrollo tecnológico como el proyecto *Molto de la UE*, el sistema *Apertium*, el sistema de código abierto *Moses* (Parra 2018) , entre otros. Sin embargo, si en el mundo existen más de siete mil lenguas, el número actual de lenguas que puedan tener sistemas de traducción todavía es mínimo. Además, muchas de las lenguas del mundo

están en peligro de extinción, y si no se crean este tipo de herramientas, posiblemente no se cree una consciencia para sistematizarlas y preservarlas a través del tiempo.

2. Marco teórico

Las nuevas tecnologías de la información y la comunicación son relativamente nuevas en su aplicación en los procesos de mediación educativa, pero su impacto es de gran importancia, ya que los nuevos paradigmas exigen nuevas formas de afrontar los procesos sociales e individuales (Domínguez Sánchez-Pinilla 2003). En cuanto a la traducción automática, existen muchas posibilidades para orientar la forma en que se puede llevar a cabo el desarrollo de una aplicación. Al realizar cualquier tipo de traducción, ya sea por medio de la intervención humana o de una computadora, el objetivo de la traducción es convertir la información de un idioma de origen y restaurarla a un idioma de destino. Hoy en día, se podría decir que existen principalmente tres tipos de traducción automática en el mercado: la que utiliza información lingüística (comúnmente conocida como RBMT por sus siglas en inglés: *Rule Based Machine Translation*), la estadística (o SMT por sus siglas en inglés: *Statistical Machine Translation*) y el basado en redes neuronales (o NMT por sus siglas en inglés: *Neural Machine Translation*). A estos tres paradigmas principales habría que sumar otras iniciativas con menor proyección, como la propuesta por investigadores de la Universidad de Gotemburgo (Suecia) basada en el concepto de *interlingua* y denominada *Grammatical Framework*.

Para el desarrollo del software traductor español-embera katío se han seguido dos etapas de prueba*: el modelo de traducción por transferencia* y *el modelo de traducción estadística* para saber cuál es más efectivo. En comparación, la traducción automática basada en reglas por transferencia permite mucha calidad en consistencia y tiene la posibilidad de adaptar la traducción al contexto gramatical; además, utiliza reglas gramaticales de oraciones para la alineación de las palabras (Hernández 2002: 109). No obstante, es difícil manejar con traducción por transferencia algunas reglas gramaticales. En la validación de las traducciones, la utilización de un modelo de traducción automático ha presentado dificultades para la lengua embera katío ya que la estructura gramatical de la lengua es compleja, como lo es por ejemplo, la utilización de marcas de caso, por ser una lengua ergativa: agente (-a), paciente (-ta); otras marcas como beneficiario (-ita), instrumental (-ba) entre muchas otras, como se puede observar en los siguiente ejemplos:

(1) m**Ʉ-a** mejaca kâga b**Ʉ**-a do-Ø-ra
1PERS.SING–ERG mucha AUX –DEC agua-NO AG-TOP
«Yo quiero mucha agua»

(2) bʉ-**a** kima-ta ero bʉ-a
2PERS.SING-ERG esposa-NO AG tiene AUX –DEC
«Tú tienes una esposa»

A nivel sintáctico (la lengua embera es del tipo SOV) las entidades pueden cambiar de posición gramatical, ya que la función es dada por las marcas gramaticales, es decir, se sabrá quién es el agente (el que realiza la acción) y quién el paciente (el que la recibe) por la marca de agente o paciente respectiva. En español estas funciones están dadas por la posición de las entidades en la oración (el español es del tipo SOV), por lo que los siguientes ejemplos también son válidos tanto en sintaxis como en semántica para la lengua embera katío:

(3) mʉ-**a** bʉ-a do-Ø-ra mejaca kâga
1PERS.SING–ERG AUX – DEC agua-NO AG-TOP mucha quiero
«Yo quiero mucha agua»

(4) bʉ-**a** bʉ-a ero kima-ta
2PERS.SING-ERG AUX –DEC tienes esposa-NO AG
«Tú tienes una esposa»

La traducción para (2) y (3) en el traductor por trasferencia es la siguiente:

(5) mʉ-**a** bʉ-a do-Ø-ra mejaca kâga
1PERS.SING–ERG AUX – DEC agua-NO AG-TOP mucha quiero
«Yo agua mucha quiero»

(6) bʉ-**a** bʉ-a ero kima-ta
2PERS.SING-ERG AUX –DEC tienes esposa-NO AG
«Tú esposa tienes»

Como se puede observar en las traducciones de (6) y (7) la transferencia que se hacen del embera katío pueden diferir mucho en la sintaxis o la semántica de lengua castellana. Por el contrario, el modelo de traducción estadística no se basa en reglas sino por las estadísticas que se hacen entre las traducciones entre dos lenguas (Hernández 2002: 109). Debido a que la lengua embera katío tiene una estructura gramatical compleja, desarrollar un traductor automático estadístico ha sido una tarea difícil ya que este modelo no toma en cuenta la gramática para realizar la traducción: este modelo suele usarse más para lenguas cuya estructura gramatical es similar. Sin embargo, este modelo ha demostrado ser más efectivo en las traducciones del español al embera katío que el

modelo por transferencia, ya que son los encargados de las traducciones quienes las van validando.

De esta manera, es importante conocer la morfosintaxis y la semántica de las dos lenguas. A continuación se muestran unos tipos de oraciones para poder identificar la morfosintaxis de la lengua embera katío (Llerena 2018: 76). La presentación de las predicaciones declarativas afirmativas simples se realiza en una tipología de oraciones monovalentes, bivalentes y trivalentes, según el número de participantes (actantes) que el núcleo predicativo requiere obligatoriamente para que se dé la relación proposicional. Esta muestra es de las oraciones monovalentes (el núcleo de la oración es un nombre, no tiene un verbo núcleo como en español).

2.1. Monovalentes

2.1.1. Oraciones de predicado nominal atributivo

Un predicado nominal es aquel cuyo núcleo o centro de predicación es una palabra que se considera nominal por las marcas que recibe, como las de caso, o una base lexical adjetival nominalizada. Las oraciones de predicado nominal están formadas mínimamente por dos nominales, siendo el primer término el que tiene la función de sujeto y el segundo el que tiene la función de predicado. En embera katío se han establecido tres clases de oraciones de predicado nominal con otro constituyente que en este caso expresa la identificación.

Este tipo de oraciones se representan con el esquema **SN+SN.**

Ejemplos:
(7) /kãɯ wẽra mɯ papa/
{kãɯ wẽra mɯ papa}
//deic.dist./mujer/1 p.s./madre//
«Esa mujer es mi mamá»

(8) /karagabira dai zeze/
{karagabi-ra dai zeze}
//Dios Karagabí-top./p.pl.pos./padre//
«Karagabí es nuestro padre»

(9) /mɯ trɯra ernesto/
{mɯ trɯ-ra ernesto}
//1 p.s.pos./nombre-top./Ernesto//
«Mi nombre (es) Ernesto»

2.1.2. Oraciones posesivas

Las oraciones de genitivo o de la posesión nos indican lo que una entidad posee o lo que le pertenece. El núcleo predicativo es un nominal con la marca casual de locativo. La presentación de estas oraciones en embera tiene la particularidad de no poseer cópula. Este tipo de oraciones se representan con el esquema: **SN+SN-gen.**

(10) /kaɯ mitʃi mɯde/
{kaɯ mitʃi mɯ-de}
//deic.dist./gato/1 p.s.-gen.//
«Ese gato es mío»

(179) /maɯ dera ernestode/
{maɯ de-ra ernesto-de}
//deic.dist./casa-top./Ernesto-gen.//
«En cuanto a la casa, es de Ernesto»

3. Metodología

Los seres humanos por naturaleza son sociables y en este proceso tienen la necesidad de comunicarse entre sí. En el proceso de comunicación hay un intercambio de información, que es procesada mentalmente y cuyo fin es codificar el mensaje. En la actualidad existen muchas culturas por todo el planeta, y el contacto entre culturas y lenguas hace posible que el hombre conozca otros pensamientos e identidades que tienen otras personas. Con la llegada de las nuevas tecnologías de la información y la comunicación, las distancias se acortan; el intercambio de información entre individuos, aplicaciones como chat, videos, podcasts, correos electrónicos entre otros, son útiles a la hora de comunicarse.

Esta investigación se ha realizado con comunidades embera katío del municipio de Tierralta (Córdoba, Colombia) y en las regiones aledañas al Nudo de Paramillo. Estas comunidades han tenido procesos etnoeducativos en las escuelas ubicadas en los asentamientos de los ríos Esmeralda, Verde y Sinú. Se ha desarrollado un sistema de escritura y en la práctica han producido textos como periódicos, cuadernillos y folletos en su idioma. Sin embargo, estos procesos han sido muy lentos; además, existen muchas dificultades para imprimir y difundir materiales auténticos en las escuelas y para el público en general. Muchas personas de estas comunidades ven el idioma español y las nuevas tecnologías como una «panacea», ya que tienen mayor contacto con los «colonos» y los habitantes no

indígenas de la zona. Asimismo, requieren el aprendizaje del español en escuelas dentro de una pedagogía intercultural, que en la zona son bastantes, pero no existe un PEI (Proyecto Educativo Institucional) acorde a sus necesidades, por lo que existe un bajo rendimiento académico, principalmente en la asignatura de lengua española. En la zona también hay muchas personas no nativas, que tienen contacto diario con estas comunidades y que desconocen por completo el idioma embera katío, por lo que requieren traductores o personas que conozcan el idioma.

Este proyecto de investigación se desarrolla con base en la Metodología de Investigación en Ciencias del Diseño (DSRM) (Hevner et al. 2004: 75) como marco para el uso de las TIC en entornos educativos. DSRM es un método utilizado en investigación para respaldar el diseño y la construcción de todo tipo de artefactos. Según Hevner (2004: 75), este tipo de métodos resuelven problemas que son difíciles de resolver porque son incompletos, contradictorios o sus requisitos pueden cambiar con el tiempo y muchas veces son difíciles de evaluar (Oscina et al. 2020: 4). Esta metodología sugiere que existen diferentes métodos o lineamientos para el desarrollo de productos tanto tangibles como intangibles que el diseñador adapta de acuerdo a las necesidades y propósito de su proyecto. En algunos campos, como la ingeniería de software, tienen ofertas específicas (Wieringa 2014: 41; Halonen 2012: 22). El pragmatismo de gran parte de la ingeniería del software sugiere una preferencia por el método de investigación constructivo o aplicado sobre la investigación pura. La metodología de la ciencia del diseño incluye tres ciclos: ciclo de relevancia, ciclo de rigor y ciclo de diseño. La herramienta principal de estos ciclos es investigar y buscar información útil para crear un artefacto en contexto. El ciclo de relevancia consiste en las necesidades del mercado y el contexto en el que se utilizará el producto para comprender mejor los requisitos y demandas del proyecto. Un ciclo de rigor es la búsqueda constante de información relacionada con las necesidades identificadas, incluidas las soluciones anteriores, es decir, la búsqueda de referencias y la información técnica necesaria para su implementación. En el ciclo de diseño se construyen y evalúan posibles soluciones al problema propuesto utilizando diversos medios para confirmar su contribución.

Para el desarrollo del software traductor español-embera katío se han seguido dos etapas de prueba: el *modelo de traducción por transferencia* y el *modelo de traducción estadística*. Este último casi no ha sido utilizado en proyectos con lenguas indígenas; mayormente se ha utilizado con lenguas SAE (Standard Average European). Este modelo es la estrategia que actualmente estamos desarrollando hacia una fase alfa, es decir, que esté listo para que los prueben los usuarios y se puedan utilizar de manera pública para contrarrestar la pérdida del idioma embera katío en Colombia.

4. Primera fase: Modelo de traducción automática basado en reglas por transferencia

El principio en el que se basó este software traductor se refiere a trasladar el proceso que realiza la mente para traducir, a condiciones que la computadora pueda interpretar; para este caso se utilizó el lenguaje de programación ActionScrip 3.0. La primera de las reglas consiste en tomar las palabras que se digitan en el cuadro de texto y separarlas mediante un método, que detecta cada vez que se hace un espacio, luego toma cada palabra separada por espacio y la almacena en un arreglo; es así como tenemos, por ejemplo, la oración: «mañana tengo partido de fútbol» = «mañana», «tengo», «partido», «de», «fútbol», de esta forma el programa comienza a detectar cuáles han sido las palabras que se han escrito para traducir. Aunado a esto, previamente se ha condicionado al usuario para que en el cuadro de texto no se puedan escribir mayúsculas, comas y puntos. Luego de esto, el software elimina las palabras que no son necesarias y no tienen ninguna traducción a la otra lengua, sin afectar la traducción; por ejemplo, los artículos *el, la, los, las, un, unas, unos, unas* que son de la lengua castellana y no tienen traducción directa en la lengua embera katío, por lo que solo se debe traducir las palabras que sean necesarias.

Un factor que determina el éxito de un traductor es la confiabilidad de su base de datos y la efectividad del programa para buscar rápidamente palabras en ella. Por ello, la base de datos de este software fue construida en XML, ya que permite una fácil construcción y acceso a sus nodos; Cada nodo de la base de datos tiene cuatro atributos importantes, la palabra en español, su traducción en embera katío, el tipo de palabra, y un auxiliar en el caso de plurales, pronombres, sustantivo, pronombre posesivo, etc. El programa cuenta con más de 5000 nodos que tienen la misma estructura. De esta forma, al acceder a un nodo, se está accediendo a toda la información necesaria para su posterior procesamiento. Lo importante de la traducción automática es asegurarse de que la computadora realice los pasos necesarios para que, al realizar el cambio de palabras, sea lo más preciso posible. Es por esto por lo que la base de datos que se creó para este traductor automático embera katío tiene aproximadamente 5000 palabras, pero tener una base de datos tan grande puede hacer que el programa tarde mucho en encontrar las palabras y reemplazarlas. Asimismo, la lengua embera katío es una lengua aglutinante con muchos morfemas de casos muy diferente a la de la lengua castellana (Llerena 2018: 15-16), lo que en la práctica puede llevar a que se presenten muchas dificultades al hacer que las traducciones sean lo más fieles al original. A continuación, una imagen de este traductor en fase beta, es decir, la primera versión del software:

Imagen 1. Traductor por transferencia español-embera katío. Fuente propia.

5. Segunda fase: modelo de traducción automática estadística CARINA

Como alternativa a los costosos procesos de desarrollo de traductores automáticos basados en reglas, se pueden utilizar los llamados *traductores automáticos estadísticos* (Parra 2018: 22). Su principal ventaja sobre los basados en reglas es que necesitan datos para ser entrenados. En concreto, necesitan un *corpus* monolingüe del idioma de destino lo más grande posible y otro paralelo con traducciones entre el idioma de origen y el de destino. Estos sistemas de traducción automática constan de tres componentes principales: *el modelo de lenguaje, el modelo de traducción* y *el decodificador*. El modelo de lenguaje se encarga de calcular la probabilidad de que una oración en el idioma de destino sea correcta. Se encarga de la fluidez de la traducción y de formar un *corpus* monolingüe de la lengua meta que se utilice lo más grande posible. El modelo de traducción se encarga de establecer la correspondencia entre el idioma de origen y el de destino y se entrena mediante un *corpus* alineado a nivel de oración. Durante esta fase de entrenamiento, el sistema estima la probabilidad de una traducción a partir de las traducciones que aparecen en el *corpus* de entrenamiento. Finalmente, el decodificador se encarga de buscar entre todas las posibles traducciones la más probable en cada caso. Así, dado un modelo de lenguaje y un modelo de traducción, crea todas las traducciones posibles y propone la más probable.

Gracias a los avances en la investigación, estos sistemas, que inicialmente producían traducciones muy malas, lograron resultados aceptables y entraron en la vida de los traductores profesionales hace unos años. Estos avances incluyen mejoras en todos los componentes de un motor de traducción automática estadística, desde la forma de realizar alineamientos suboracionales para mejorar el modelo de traducción, hasta la

incorporación de información lingüística como parte del entrenamiento, o la incorporación de otras técnicas de procesamiento. Obviamente, la calidad de estos motores depende de varios factores, como el par de idiomas utilizado, la calidad del *corpus* de entrenamiento, el campo, etc.

Con los primeros intentos de utilizar Inteligencia Artificial (IA) en entornos profesionales surgieron iniciativas que buscaban acercar la traducción automática estadística a traductores o usuarios fuera del mundo académico o de los círculos de desarrollo de motores de Inteligencia Artificial. Así nacieron proyectos como *Moses for mere mortals*, que buscaba acercar la traducción estadística a cualquier persona interesada en ella, y *Moses for localization (m4loc)*, que buscaba adaptar la traducción automática estadística para el sector de la localización. En los últimos años también han surgido soluciones comerciales como *Slate*, que pretenden facilitar a los traductores el acceso a estas tecnologías y permitirles entrenar sus propios motores de traducción automática en su ordenador (Parra 2018: 20). Otras propuestas, aunque basadas en la nube, son las de *KantanMT*, una startup irlandesa, o la estadounidense *Lilt*. Esta última también tiene la particularidad de ofrecer una función llamada «traducción automática interactiva», que permite al usuario seleccionar la siguiente palabra en la traducción y, dependiendo de la palabra seleccionada, cambiar la propuesta de traducción. Y si miramos los resultados de los proyectos de investigación europeos encontramos *MateCAT*, una herramienta de traducción asistida en la nube que incorpora traducción automática.

El grupo de Investigación EduTLan (grupo de investigación de la Facultad de Educación de la Universidad de Córdoba) ha estado investigando sobre estos sistemas de Inteligencia Artificial (traductores automáticos estadísticos, traductores automáticos basados en reglas, proyectos como *Moses for mere mortals*, *KantanMT*, *Lilt MateCAT*, entre otros) con el fin de determinar aquellos que pueden ser útiles a la hora de tomar decisiones en cuanto al desarrollo del software traductor. Para el desarrollo del modelo estadístico de traducción automática para el idioma embera katío se está utilizando CARINA, la cual es una arquitectura metacognitiva estadística para agentes cognitivos que tienen un perfil de conocimiento algorítmico, que tiene el estado local de una función cognitiva en forma de estados algorítmicos locales (Caro et al. 2019). La representación semántica del conocimiento es fundamental para lograr un mecanismo que proteja la información que un agente cognitivo recoge del mundo y de sí mismo; esta representación semántica en la arquitectura CARINA es necesaria para identificar la información que se encuentra en un perfil de conocimiento algorítmico y así poder acceder a ella para la localización de fallas de razonamiento. Posteriormente, se presenta un perfil de conocimiento algorítmico en forma de creencia que es la unidad mínima de conocimiento declarativo en la memoria semántica de CARINA. Para la traducción de oraciones del español al idioma embera katío, se alimenta el sistema CARINA con oraciones simples y complejas en ambos

idiomas y se valida su efectividad. A continuación, una imagen de este traductor en fase beta, es decir, la primera versión del software:

Imagen 2. Traductor estadístico español-embera katío. Fuente propia.

6. Validación del software de traducción

En general, los traductores para idiomas tradicionales como el inglés o el francés se encuentran virtualmente (en línea) o en procesadores más pequeños. Estos traductores tienen cierta fiabilidad, pero hasta el momento no hay ninguno que sea cien por cien fiable, ya que hay muchos problemas con la semántica y sintaxis de los idiomas a la hora de realizar las traducciones. En el caso de las lenguas indígenas, existen pocos traductores de estas lenguas, ya que muchas no cuentan con sistemas de escritura, o si los tienen, varían según la comunidad o las propuestas de los investigadores. Así, los indígenas embera katío que han cursado alguna carrera académica en la Universidad de Córdoba ayudan para este fin. Ya que el modelo de traducción estadístico requiere añadir información de oraciones en español y embera constantemente, los indígenas que ayudan en este proceso validan continuamente las traducciones. Por ejemplo, esta es una traducción de un texto traducido con Iván Domicó (Llerena 2018: 135), un estudiante indígena de la carrera de Licenciatura en Sociales de la Universidad de Córdoba, de la etnia embera katío de la comunidad de Pãwarando, quien ha venido colaborando como traductor hace 6 años.

Imagen 3. Ayudante de traducción embera katío Iván Domicó, su esposa y el profesor Ernesto Llerena. Fuente propia.

DESCRIPCIÓN DE ANIMALES

opoga
//iguana//

1. opoga do kida bema.
{opoga do kida bema.}
//iguana/río/orilla/proceder//
«La iguana procede de la orilla del río»

2. opoga ne tuku komia
{opoga ne tuku komia}
//iguana/ref.def./cogollo/comedor//
«La iguana es comedora de cogollo»

3. iyi kakuara pãwara pãwara chʉ bʉa
{iyi kakua-ra pãwara pãwara chʉ bʉ-a}
//3 p.s.pos./cuerpo-top./azuloso/est./aux.ser-decl.//
«Su cuerpo es azuloso»

4. iyi dru kõgo kõgo drasoa chʉ bʉa
{iyi dru kõgo kõgo drasoa chʉ bʉ-a}

//3 p.s.pos./cola/rayada en círculos/larga/est./aux.ser-decl.//
«Su cola es rayada en círculos»

5. iyi jɯwa jîwini juesoma drasoa chɯ bɯa
{iyi jɯwa jîwini juesoma drasoa chɯ bɯ-a}
//3 p.s.pos./pata/dedos/cinco/gruesos/est./aux.ser-decl.//
«Su pata tiene cinco dedos largos»

6. êbêra opoga ko bada
{êbêra opoga ko bada}
//indígenas/iguana/comer/hab.pl.//
«Los indígenas comen la iguana»

7. kûrama wia bada
{kûrama wia bada}
//ahumar/cocinar/hab.pl.//
«La cocinan ahumada»

8. pata eda ko bada
{pata eda ko bada}
//plátano/soc./comer/hab.//
«Se la comen con plátano»

9. neta torro pichi eda opoga warra
{neta torro pichi eda opoga warra}
//arroz/soc./iguana/sabrosa//
«La iguana es sabrosa con arroz»

7. Conclusiones

El trabajo relacionado con la traducción automática, con población indígena, es nuevo en Colombia. La utilización de un modelo de traducción automático por transferencia presenta dificultades para la lengua embera katío, ya que la estructura gramatical de la lengua es compleja, como lo es, por ejemplo, la utilización de muchas marcas como las de caso, y variaciones en el orden de las palabras en la oración simple y compleja (Llerena, 2018). De igual manera, la traducción de las demás lenguas indígenas en Colombia puede presentar dificultades para este tipo de traducción por transferencia debido a la morfosintaxis de estas lenguas. A pesar de que el modelo estadístico requiere almacenar mucha información de traducción de oraciones y textos de las dos lenguas, lo

que conlleva a implementar mucho tiempo en esta labor, es muy efectivo, ya que se pueden ir mejorando las traducciones continuamente.

Mirando la situación actual de las etnias del país, con los aportes de las tecnologías de la información, podemos ayudar a preservar culturas que siempre han tratado de mantenerse por generaciones, pero que, por las diferentes problemáticas sociales que se presentan en Colombia, estas comunidades se han ido rezagando y han sido olvidadas por la sociedad. Este proyecto no solo puede contribuir a la preservación, sino que también puede servir como modelo para desarrollar nuevas formas o aplicaciones de comunicación para poner en valor las otras sesenta y cuatro lenguas indígenas y dos criollas en Colombia, y las más de cinco mil lenguas del mundo que no cuentan con este recurso.

BIBLIOGRAFÍA

Barrett, Mandy, et al. «Using Artificial Intelligence to Enhance Educational Opportunities and Student Services in Higher Education.» *Inquiry: The Journal of the Virginia Community Colleges* 22.1 (2019): 1–11. Web. 15 de abril 2023.

Becerra Cortés, Yunuén Esperanza. «Estudiantes indígenas y los usos y apropiación de las tecnologías de información y comunicación.» *Paakat* 2.3 (2012): s. p. Web. 5 de septiembre 2013.

Caro, Manuel Fernando, et al. «The Carina metacognitive architecture.» *IJCINI* 13.4 (2019): 71–90. Web. 15 de abril 2023.

Congreso de Colombia. *Ley para la protección de las lenguas nativas. Ley 1381*. Bogotá, 2010. Impreso.

Domínguez Sánchez-Pinilla, Mario. «Las tecnologías de la información y la comunicación: sus opciones, sus limitaciones y sus efectos en la enseñanza.» *Nómadas* 8 (2003): s. p. Web. 15 de abril 2023.

Google. *Traductor de Google*. 2006. Web. 15 de abril 2023.

Halonen, Nillo. «Product Life-Cycle Disposition Model-Disposition Conceptualising for Design Science.» Tesis de Maestría. Universidad Tecnológica de Tampere, 2012. Web. 15 de abril 2023.

Hernández, David, *et al*. *Traductor Español a Náhuatl.* Web. 27 Oct. 2015.

Hernández, Pilar. «En torno a la traducción automática.» *Cervantes* 2 (2002): 101–117. Web. 15 de abril 2023.

Hevner, Alan, et al. «Design Science in Information Systems Research.» *MIS Quaterly* 28.1 (2004): 75–105. Web. 15 de abril 2023.

Llerena, Ernesto. *Fundamentos gramaticales de la lengua embera.* Montería: Editorial Zenú, 2018. Impreso.

Llerena, Rito. «Sintaxis de la predicación de la lengua ēpēra oriental del Alto Andágueda.» *Bulletin de l'Institut Français d'Études Andines* 23.3 (1994): 437–462. Impreso.

Lu, Joyce, et al. *Artificial Intelligence (AI) and Education. FOCUS: Congressional Research Service*. Web. I de agosto 2018

Ministerio de Cultura. *Ley para la Protección de las Lenguas Originarias*. *Programa para la Protección a la Diversidad Etnolinguística (PPDE)*. Colombia, 2010. Impreso.

Moses for localization (m4loc). *Sistema de código abierto para traducción Public MT*. Web. 2010.

Moses. *Moses for mere mortals. Public MT*. Web. 2015.

Moses. *Sistema de código abierto para traducción. Public MT.* Web 2022.

Molto Molto Project. Web. 2023

Oscina, Luis, et al. «Proceso de Design Science Research aplicado a la Construcción de una Ontología de Testing de Software como Artefacto.» *Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza* 5.1 (2020): 1–22. Web. 15 de abril 2023.

Parra, Carla. «¿Cómo ha evolucionado la traducción automática en los últimos años?» *La linterna del traductor* 16 (2018): s. p. Web. 15 de abril 2023.

Rojas, Tulio, et al. «Building a Nasa Yuwe Language Corpus and Tagging with a Metaheuristic Approach.» *Computación y sistemas* 22.3 (2018): 881–895. Web. 15 de abril 2023.

Simón, Fray. *Noticias Historiales de las Conquistas de Tierra Firme en las Indias Occidentales*. Bogotá Biblioteca de Autores Colombianos, 1953. Impreso.

TraductorPro.com.Web. 2023.

UNESCO. *Atlas de las lenguas del mundo*. Web. 2022.

Werner Cantor, Erik. *Ni Aniquilados, Ni Vencidos: Los embera y la gente negra del Atrato Bajo el Dominio español. Siglo XVIII*. Bogotá: ICANH, 2000. Impreso.

Wieringa, Roel. *Design Science Methodology for Information Systems and Software Engineering*. Berlin: Springer, 2014. Print.

Fecha de recepción: 14 de noviembre de 2022
Fecha de aceptación: 8 de abril de 2023