Artículo de investigación

Recopilación de corpus paralelo español-guaraní y experimentos iniciales con traductor automático estadístico

Initial Parallel Corpus Creation and Statistical Machine Translation Experiments for Spanish Guarani pair of Languages

Revista sobre estudios e investigaciones del saber académico

Aldo Andrés Álvarez López[1]
[1]Universidad Nacional de Itapúa, Dirección de Investigación y Ambiente. Facultad de Ingeniería. Encarnación, Paraguay. aldo.alvarez@fiuni.edu.py
https://orcid.org/0000-0002-0443-9198

Resumen

En este artículo se presenta el trabajo realizado para recolectar conjuntos de oraciones en español y guaraní a fin de crear un corpus bilingüe que servirá como base para la creación de tecnología lingüística relacionada con el par de idiomas. En este caso, se hace foco en la traducción automática del español al guaraní. El guaraní es una lengua que carece, en gran medida, de recursos digitales. Esto impide que la misma prospere en cuanto a tecnología se refiere. Para la generación del corpus se ha hecho uso de materiales digitales disponibles en la nube. Así también, se ha utilizado una plataforma web denominada Guampa con el objetivo de generar nuevas frases de forma colaborativa. Se presentan datos estadísticos del corpus generado y experimentos iniciales con Moses y su plataforma para la Traducción Automática Estadística (*SMT*, del inglés, *Statistical Machine Translation*). Los resultados pretenden servir de punto de partida para futuros experimentos en el área.

Palabras claves: Corpus paralelo. Corpus bilingüe. Traducción automática estadística. Guaraní.

Abstract

This paper introduces the work that has been done to collect sentences in Spanish and Guaraní to create a bilingual corpus. This corpus might serve as a baseline for the creation of linguistic technology related to the pair of languages. In this article, the focus is on machine translation from Spanish to Guaraní. Guaraní is an under-resourced language that suffers from digital resource insufficiency. This prevents the language from thriving in terms of technology development. To generate the bilingual corpus, digital resources available on the cloud have been used. Furthermore, a web platform called *Guampa* has been employed to generate new phrases collaboratively. Statistical data related to the corpus is presented along with initial experiments for Statistical Machine Translation (SMT) using Moses platform. The results serve as a starting point for future research in the area.

Keywords: Parallel corpus. Bilingual corpus. Statistical machine translation. Guaraní.

Área del conocimiento: Ingeniería y Arquitectura
Correo de Correspondencia: aldo.alvarez@fiuni.edu.py
Conflictos de Interés: El autor declara no tener conflictos de intereses.

Fecha de recepción: 13/01/2022 Fecha de Aprobación: 25/05/2022
Página Web: http://publicaciones.uni.edu.py/index.php/rseisa

Citación recomendada: Álvarez López, A. A. (2023). Recopilación de corpus paralelo español-guaraní y experimentos iniciales con traductor automático estadístico. Revista sobre estudios e investigaciones del saber académico (Encarnación), 17(17): e2023003

Introducción

La Traducción Automática Estadística (*SMT*, del inglés, *Statistical Machine Translation*) requiere de corpus bilingües extensos, también denominados corpus paralelos, para generar resultados óptimos y confiables. Por dicha razón, las soluciones computacionales relacionadas con la traducción automática son difíciles de implementar en escenarios donde los recursos digitales son escasos. Esto sucede especialmente con idiomas nativos, con aquellos en peligro de extinción o que por razones sociales o históricas fueron relegados a un plano inferior (Gasser, 2006).

El guaraní es un idioma originalmente nativo que se ha convertido, conjuntamente con el español, en la lengua oficial del Paraguay. Es hablado por más de 12 millones de personas, nativas y no nativas, en regiones que incluyen también a Argentina, Brasil y Bolivia. Además, es enseñado en escuelas, colegios y universidades del Paraguay. Pese a esta situación de importancia, el acceso al material escrito digitalizado es limitado. Esto es debido a razones socio-históricas que han puesto al guaraní en una situación de desventaja comparado con el español.

Contar con un traductor automático y otras tecnologías relacionadas al procesamiento del lenguaje natural (*NLP*, del inglés, *Natural Language Processing*) promovería la producción de material digital y facilitaría la promoción del idioma ya que sería utilizada por profesionales, estudiantes y adeptos de la lengua. Además de estrechar la brecha comunicacional que existe entre sectores de la población donde algunos habitantes se comunican más fluidamente en guaraní que en español. Existen proyectos realizados para la creación de analizadores sintácticos y morfológicos para el Guaraní (*Apertium/Apertium-Grn*, 2018/2020) (*Morfo: Análisis y Generación Morfológica*, n.d.) y proyectos de Traducción Asistida por Computadora (*CAT*, del Inglés, *Computer Assisted Translation*) como Mainumby (Gasser, 2018). Sin embargo, se hallan limitados experimentos relacionados con *SMT*. Esto en consecuencia de la no existencia de un corpus bilingüe extenso, lo que ralentiza el avance de la Traducción Automática (*MT*, del inglés, *Machine Translation*). Por otra parte, existen otras características del idioma que suman complejidad a la hora de implementar un sistema de traducción. Morfológicamente, el guaraní es considerado como polisintético-aglutinante, lo cual refiere a que las palabras son formadas a través de la agrupación de morfemas por medio de prefijos y sufijos. Una sola palabra en guaraní podría significar una oración completa en español.

El presente proyecto se centra en la generación de un corpus bilingüe español-guaraní y en pruebas iniciales con *SMT* haciendo uso de la plataforma Moses (Koehn et al., 2007). El artículo está organizado de la siguiente manera. En la sección 2 se mencionan trabajos previos relacionados a *NLP* y *SMT*. La sección 3 menciona características específicas del idioma guaraní. La sección 4 habla de la recopilación de oraciones para el corpus bilingüe. La sección 5 y 6 mencionan pruebas experimentales y resultados. La sección 7 expone conclusiones y trabajos futuros.

Trabajos relacionados

Existen trabajos que se han realizado en torno al guaraní en cuanto a *NLP* se refiere. Estos proyectos incluyen temas referentes a recolección de corpus, análisis morfológicos y sintácticos, segmentación de palabras, etiquetado gramatical o *Part of Speech Tagging* y traducción asistida por computadoras. A continuación, se presentan los más relevantes.

Abdelali et al. (Milagros et al., 2006) introdujeron una serie de procedimientos a fin de colectar corpus paralelos para idiomas con recursos limitados que incluía al guaraní como parte del caso de estudio. Según el proyecto, 250.000 oraciones bilingües, guaraní-inglés, fueron colectadas. Sin embargo, los datos no se encuentran públicamente disponibles. En la misma línea, Rudnick et al. (Rudnick et al., 2014) desarrollaron Guampa, una aplicación web para recolectar oraciones bilingües español-guaraní de manera colaborativa. El proyecto es accesible públicamente y se ha utilizado en esta investigación para obtener oraciones de forma colectiva en conjunto con estudiantes del idioma guaraní.

En cuanto a herramientas lingüísticas para el guaraní podemos citar a MORFO (*Morfo: Análisis y Generación Morfológica*, n.d.), proyecto del laboratorio de Tecnología del Lenguaje Humano y Democratización de la Información (*HLTDI*, del Inglés, *Human Language Technology and the Democratization of Information*) de Indiana University. MORFO es un programa que permite el análisis morfológico de varios idiomas nativos y de

escasos recursos, entre ellos el guaraní. La aplicación permite obtener información detallada de una palabra dada, entre las cuales podemos mencionar el número, la raíz de la palabra, el tiempo verbal, el caso, entre otros. Por otro lado, Apertium (*Apertium/Apertium-Grn*, 2018/2020), una plataforma de código abierto para traducciones, incluye un paquete de herramientas para el análisis, la segmentación, la generación de palabras y el etiquetado gramatical en Guaraní. En un trabajo futuro, se pretende utilizar estas herramientas para la mejora del sistema de traducción. Por otro lado, Maldonado et al. (Maldonado et al., 2016) presentaron Eñe'e, un sistema de reconocimiento del habla en guaraní. En dicho proyecto, se ha logrado recolectar 1000 oraciones, con 678 palabras únicas. Con su sistema lograron reconocer alrededor del 90% de las palabras utilizadas como parte de la validación de la herramienta.

Por último, a lo que refiere a Traducción Automática (*MT*, del Inglés, *Machine Translation*), Gasser, M. (Gasser, 2018) introdujo un ayudante para la traducción Castellano-Guaraní denominado Mainumby. El sistema realiza traducciones haciendo uso de léxicos bilingües, reglas de transformación, un analizador morfológico para español y un generador morfológico para el guaraní. Hasta el momento, Mainumby es uno de los sistemas más completos para la traducción del español al Guaraní. Sin embargo, no puede hacer uso de técnicas más complejas como *SMT* para generar traducciones por la falta de un corpus paralelo cuyo tamaño y calidad permita obtener mejores resultados. La creación y extensión de un corpus bilingüe será de relevancia para que esta herramienta pueda mejorar las traducciones.

Características del Guaraní.

El guaraní es uno de los idiomas oficiales del Paraguay conjuntamente con el español. Éste posee la particularidad de que entre las lenguas nativas de América es la única utilizada mayormente por no nativos (mestizos o criollos). Pertenece a la familia Tupí-guaraní y es hablada por más del 80% de la población paraguaya.

El alfabeto Guaraní hace uso de 33 fonemas, representados por 12 vocales y 21 consonantes clasificadas en dos categorías: orales o nasales. Las palabras también son clasificadas de esta manera dependiendo de las vocales y consonantes que las conforman. Entre las consonantes, podemos encontrar dígrafos como la mb, nd, ng, nt,rr que representan cada uno un fonema determinado. También se hace uso de la tilde (~) para representar el sonido nasal de las vocales y el de las letras ñ y **g̃**, aunque esta última a veces es reemplazada por la ^***g***, ya que no se encuentra la representación de este carácter en el estándar UNICODE. Otro fonema especial es el silencio o pausa gutural que está representado por el apóstrofe o comilla simple, más comúnmente conocido como pusó en el idioma Guaraní (*Guarani Language and the Guarani Indian Tribe (Avañe'e, Jopará, Chiriguano, Mbyá)*, n.d.).

Sintácticamente, una oración en guaraní sigue el orden sujeto-verbo-predicado o SVO (del inglés, *Subject-Verb-Object*) al igual que el español. Morfológicamente, el Guaraní se considera polisintético-aglutinante (Verón & Marae'y, 2018) en el que a la palabra raíz o lexema se agregan prefijos y sufijos para conformar una nueva palabra, o modificar el significado de la misma. En muchas ocasiones una sola palabra guaraní representa toda una oración en castellano. Además, en muchas ocasiones la versión en guaraní de ciertas oraciones no es una traducción directa de la versión en español. En la Tabla 1 se puede observar un ejemplo de oración en español y su traducción al guaraní, extraído de los datos recolectados. En este ejemplo podemos observar algunas de las características sobre el idioma guaraní mencionadas más arriba. El idioma no posee género ni artículos definidos y los adjetivos no existen como categoría separada salvo los calificativos.

Tabla 1.
Ejemplo de una oración en Castellano y su correspondiente traducción al guaraní. Fragmento extraído de la página de noticias de la Secretaría de Políticas Lingüísticas (SPL) del Paraguay.

Español	Guaraní
El taller de normalización del uso del guaraní en instituciones públicas, realizado este martes por la Secretaría de Políticas Lingüísticas (SPL) en el Departamento de Itapúa, tuvo una activa participación de las autoridades locales y de los diferentes ámbitos de la sociedad.	Guarani ñe'ẽ jeporumeme tetã remimoĩmby rembiapópe, ombosako'íva Paraguái Ñe'ẽnguéra Sãmbyhyha (PÑS), tavusu Itapúa pe, ombyaty heta tapichápe, heta temimoĩmbýgui ha tavusu tuichakue.

Recopilación de oraciones bilingües para la generación del corpus

Para poder realizar experimentos relacionados con *SMT* es necesario contar con un corpus paralelo que contenga un número significativo de oraciones. Para poder recolectar un corpus bilingüe se ha recurrido a recursos disponibles en la web. Así también se ha organizado una actividad denominada Hackathon Lingüístico, para recolectar corpus de forma colaborativa con estudiantes del idioma.

En la Tabla 2 se pueden ver datos relacionados a las oraciones recolectadas. A continuación, se describen los diferentes recursos utilizados para construir el corpus utilizado en los experimentos. Probablemente, existen otros materiales que no se han incluido como parte del corpus para esta investigación. Las razones posibles son que dichos materiales no se encuentran apropiadamente digitalizados para su uso como parte del conjunto de datos o no eran accesibles públicamente a través de internet. En caso de conocer acerca de materiales que puedan ser incorporados al corpus, por favor ponerse en contacto con el autor del artículo.

Tabla 2.

Datos estadísticos acerca de las fuentes consultadas para obtener oraciones bilingües y monolingües.

Fuente	Total de oraciones bilingües	Promedio de palabras por oración.		Español (es)		Guaraní (gn)	
		es	gn	Total palabras	Vocabulario	Total palabras	Vocabulario
Seminario de lenguas	1.967	23	15	45.482	7.078	30.045	8.076
SPL - Noticias	4.562	29	20	131.527	8.681	91.157	10.302
Mainumby	1.600	20	14	14.389	3.310	10.116	3.322
Biblia	22.818	23	17	516.614	24.828	386.539	31.074
Seminario de lenguas - Libro 3.	1.620	15	10	3.305	1.466	2.885	1.519
Hackathon Lingüístico	800	6	5	12.100	811	3.467	937
Vikipetа	14.686	-	14	-	-	273.225	42.353

Nota: Se puede observar que en la versión guaraní de las oraciones se necesitan menos palabras. Sin embargo, el vocabulario (palabras únicas) en guaraní es mayor que en español. Esto último se podría considerar un indicativo de la naturaleza aglutinante del idioma.

Texto religioso.

La Biblia es el libro que más ha sido traducido a lo largo de la historia. Existe una versión también para el guaraní. En el repositorio del laboratorio *HLTDI* se encuentra un corpus paralelo extraído de la biblia (*Hltdi/Bitext*, n.d.). Aunque el texto religioso posee un gran número de oraciones, el vocabulario es bastante específico y el dominio del lenguaje es limitado. No obstante, debido a la escasez de material digital disponible, sería una desventaja no hacer uso de estas traducciones. El total de oraciones obtenidas de la biblia es de 22.818.

Noticias publicadas en la página de la SPL

La Secretaría de Políticas Lingüísticas (*SPL*), desde su portal online, publica noticias o acontecimientos en Español y su correspondiente versión en Guaraní (*SPL :: Marandukuéra*, n.d.). Ésta secretaría es la entidad oficial desde el estado para promover el uso de la lengua. Gracias a un software de extracción automática de información, creado para este proyecto, se han extraído las versiones en español y sus correspondientes traducciones de las publicaciones hechas desde el año 2017 hasta inicios del 2020. Luego del preprocesamiento de los archivos como eliminación de duplicados, traducciones no existentes de una oración a otra y la alineación de las oraciones en español con su correspondiente par en Guaraní se obtuvieron un total de 4.562 oraciones.

Mainumby

Como se menciona más arriba, Mainumby (Gasser, 2018) es una herramienta para la traducción asistida por computadora. Como parte del corpus utilizado para crear sus memorias de traducción, se han utilizado traducciones de libros de autores paraguayos. Se han adjuntado estas oraciones para extender el número de oraciones del conjunto de datos. Se han obtenido 1.600 oraciones de este proyecto.

Wikipedia

Existe una versión de Wikipedia para guaraní de la cual se han extraído 14.612 oraciones monolingües (solamente en Guaraní) que servirán como base para la creación de un modelo de lenguaje necesario para posteriormente poder estimar un traductor estadístico.

Hackathon Lingüístico

Como parte del proyecto, se ha llevado a cabo una actividad denominada Hackathon Lingüístico. En dicha actividad se reunieron estudiantes de la carrera de Bilingüismo Guaraní-Castellano de la Universidad Nacional de Itapúa, en Encarnación-Paraguay con el objetivo de traducir al guaraní oraciones existentes en español. Se ha utilizado una versión adaptada de la herramienta Guampa (Rudnick et al., 2014). En 4 horas de trabajo se han recolectado 800 oraciones. Como versión experimental, la actividad resultó exitosa y se pretende realizar otras ediciones para promover la generación de material digital bilingüe para el par de idiomas.

Libro: Jarojera Guarani ÑE'Ẽ

Como parte del Seminario Internacional sobre Traducción, Terminología y Lenguas Minorizadas, se ha publicado un material bilingüe (Verón & Marae'y, 2018) que contiene presentaciones de proyectos y artículos presentados en mencionado seminario. Este documento fue procesado y del mismo se han extraído 1.967 oraciones en español con sus correspondientes traducciones. También, en el mismo seminario se ha presentado el Libro 3 conteniendo vocabulario en guaraní con sus traducciones. Del mismo se han extraído 1.620 frases cortas.

Todas estas oraciones se han utilizado para construir un corpus bilingüe inicial que cuenta, hasta el momento, con 33.367 oraciones en guaraní con su correspondiente traducción al español y 14.686 oraciones en Guaraní sin traducción.

Experimentos iniciales

Para los experimentos relacionados a *SMT* se ha utilizado la herramienta Moses (*Moses - Main/HomePage*, n.d.). Moses es un sistema de traducción automática estadística que permite el entrenamiento de traductores para cualquier conjunto de idiomas siempre y cuando se cuente con el corpus paralelo correspondiente. Una vez que el sistema es entrenado, se pueden obtener traducciones correspondientes de un idioma a otro.

Para este proyecto, se ha utilizado el modelo traductor básico de Moses que está basado en traducciones por frase (*phrase-based translation*).

Tabla 3.

División del corpus en datos de entrenamiento, validación y testeo. La configuración básica de Moses requiere dividir los datos en estos 3 subconjuntos.

	Entrenamiento	Validación	Testeo
Total oraciones	26.379	4.057	2.931
% por subconjunto	79.06%	12.16%	8.78%

Preprocesamiento

Se han seguido los siguientes pasos de preprocesamiento de las oraciones bilingües:

- Normalización de los caracteres especiales y conversión a minúsculas. En Guaraní dependiendo de la fuente del material, el acento nasal es representado por una diéresis (ä), una tilde (ã) o el *acento circunflejo* (^). Para normalizar los datos, se han reemplazado tanto la diéresis como el acento circunflejo por la tilde. En el caso de la letra 'g' se ha utilizado el acento circunflejo ya que la 'g' con tilde no existe en la codificación ASCII (*American Standard Code for Information Interchange*). Posteriormente se han convertido en minúsculas todas las palabras de las oraciones.
- Eliminación de signos de puntuación. Para el mejor funcionamiento del sistema de traducción y su

entrenamiento es necesario remover algunos signos de puntuación. En este caso, se han removido todos los signos excepto los puntos (.) y comas (,) ya que los mismos pueden servir de puntos de referencia a la hora de la alineación de palabras y la traducción.
- División del corpus en conjuntos de entrenamiento, validación y testeo. El subconjunto de entrenamiento es utilizado para entrenar el traductor propiamente, es decir, son las oraciones de las cuales se van a aprender características probabilísticas a ser utilizadas para generar traducciones. Los datos de validación son utilizados para ajustar el modelo, verificar que tan buenas son las traducciones generadas y qué parámetros hay que modificar en el modelo para obtener mejores resultados. Los datos de testeo se utilizan para medir el rendimiento final del modelo, en nuestro caso, se utilizarán para cuantificar la eficacia del traductor. La división de los datos se puede observar en la Tabla 3. Se ha separado el conjunto de datos en subconjuntos de entrenamiento, validación y testeo respectivamente ya que la configuración básica de Moses requiere dividir los datos de esta manera a fin de obtener un modelo con hiperparámetros ajustados correctamente. No se ha considerado utilizar el método de validación cruzada, que usualmente se aplica para entrenar y validar modelos estadísticos cuando el número de datos de entrenamiento es reducido, ya que la configuración de Moses no permitía esta opción de manera sencilla. No obstante, dicho experimento puede realizarse como trabajo futuro y verificar si mejora los resultados obtenidos en este trabajo.

Para poder entrenar el traductor con Moses, primero hay que generar un modelo de lenguaje para asegurar la fluidez en las traducciones. Este modelo es generado haciendo uso del lenguaje al cual se quiere traducir, en nuestro caso el guaraní. Se han utilizado todas las oraciones en guaraní del conjunto de oraciones bilingües de los conjuntos de entrenamiento y validación aumentados con las oraciones monolingües sin traducción que se han colectado. Los detalles de cómo generar este modelo de lenguaje se encuentran en la documentación de Moses (*Moses - Main/HomePage*, n.d.).

Finalmente, se ha entrenado el modelo traductor siguiendo las especificaciones de Moses y los pasos correspondientes a la generación de un modelo traductor como ser la alineación de palabras, haciendo uso de *GIZA++*. *GIZA++* es un programa que se encarga de la alineación automática de palabras entre oraciones bilingües. Seguidamente se ha realizado la extracción de frases, la generación de tablas léxicas, el cómputo de un modelo de reordenamiento y el generador de traducciones propiamente. Luego del entrenamiento, se procedió al ajuste del modelo con los datos de validación para obtener mejores parámetros para la estimación de las traducciones. Una vez ajustado el modelo, se procedió a la evaluación del mismo con los datos de testeo.

Resultados

La métrica utilizada para evaluar la calidad de las traducciones obtenidas fue el puntaje *BLEU* (Papineni et al., 2002). *BLEU* es un método para evaluar traducciones automáticas y que ha sido ampliamente utilizado para medir la eficacia de traductores automáticos. Cuanto más alto sea el puntaje *BLEU*, mejor es la traducción obtenida en comparación con la traducción realizada por un humano.

Tabla 4.

Resultado de la evaluación de las traducciones obtenidas desde el traductor basado en los datos de testeo.

	Subconjunto de Testeo.			
Modelo	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Moses basado en frases	47.1	22.7	15.5	11.3

En la Tabla 4 se puede observar el resultado obtenido con el traductor base utilizando el corpus bilingüe recolectado. El puntaje de *BLEU* se ha calculado sobre los datos del subconjunto de Testeo. Los subíndices de *BLEU* indican los n-gramas que se han utilizado para medir las coincidencias entre las traducciones obtenidas por el traductor automático y las traducciones verdaderas. Podemos observar que para la métrica *BLEU-1* tenemos un puntaje de 47.1 en una escala de 100. Es un puntaje bastante bueno si consideramos que las oraciones son diversas en

cuanto al dominio del conocimiento al cual pertenecen (algunas oraciones son académicas, otras religiosas, otras coloquiales). Sin embargo, *BLEU-1* solo considera las coincidencias entre palabras individuales. A medida que aumentamos la cantidad de n-gramas vemos que el puntaje BLEU se reduce. Para *BLEU-4* se obtiene un puntaje de 11.3 lo cual indica que el orden en el que aparecen las palabras en la traducción automática no tiene gran coincidencia con la traducción real de las oraciones.

En la Tabla 5 podemos observar ejemplos de traducciones realizadas por el traductor automático y la traducción real según el corpus bilingüe. Se observa como con oraciones cortas y cuyo vocabulario es más común, si lo contrastamos con el vocabulario obtenido de la fase de entrenamiento, las traducciones tienden a ser mejores. No obstante, si el vocabulario es un poco más rebuscado o técnico, las traducciones no son ni cercanas a las verdaderas. Esto da un indicio de que más oraciones bilingües son necesarias para ampliar el vocabulario del traductor y sobre todo tratar de integrar oraciones que incluyan palabras de áreas y dominios que permitan enriquecer el conjunto de palabras.

Tabla 5.
Ejemplos de traducciones obtenidas por el traductor automático y las traducciones reales con sus correspondientes puntajes de BLEU.

Castellano	Guaraní	Traducción Automática	Bleu
será en la sala de consejo del instituto superior de educación ise de 16 : 00 a 18 : 00 horas .	ko aty oikóta instituto superior de educación ise kotýpe , 16 : 00 aravo guive 18 : 00 peve .	oikóta koty consejo del instituto superior de educación ise kotýpe , 16 : 00 aravo guive 18 : 00 peve .	81.0
una extensa lista de actividades desplegadas a nivel nacional e inclusive en el exterior fueron llevadas adelante desde la comisión integrada por varias instituciones, incluida la SPL, así como destacadas figuras de la cultura nacional.	heta mba'e porã ojejapo comisión guive tetãpy tuichakue ha tetã ambue rehe avei ko ary pukukue oñemomorãvo ñande haihára guasúpe , ko comisión , omoirũva avei pñs , oike hyepýpe heta ambue temimoĩmby ha tapicha katupyry .	peteĩ mba'e desplegadas extensa rerarysýi ha oimehína ñemomba'everã , ha upe guive oñehesa'ỹijo llevadas omoirũva'ekue ko aty omoirũ pñs hína heta temimoĩmby ha tapicha , momaranduha de la cultura nacional .	8.0

Nota: La primera oración tiene una pequeña diferencia en la traducción automática, sin embargo, es cercana a la original. Por el otro lado, en la segunda oración la traducción automática es totalmente distinta, carece de sentido si la comparamos con la original y algunas palabras no son traducidas al guaraní.

Problemas de alineación.

Oraciones a ser alineadas entre sí: **Castellano:** La municipalidad de San Lázaro y la comunidad indígena del pueblo Guaná llamada Río Apa. **Guaraní:** Tavaro San Lázaro ha ypykuéra aty táva Guana hérava Río Apa.	
Alineación obtenida	Alineación Esperada

Figura 1.
Ejemplo de alineación errónea.

Nota: A la izquierda, se puede observar el resultado de un alineamiento automático de palabras. A la derecha, se tiene el resultado esperado. Sin embargo, el alineador automático no logra emparejar correctamente las palabras con sus traducciones incluso cuando estas aparecen en la versión guaraní de la oración.

Un paso crucial durante el entrenamiento del traductor es el de la alineación de palabras. En la Figura 1 se puede observar la alineación entre una

frase en español y otra en guaraní. En este ejemplo se puede observar cómo, a pesar de que las palabras en español y sus correspondientes traducciones al Guaraní aparecen en las frases, el alineador de palabras no es capaz de ubicarlas correctamente. Esto puede deberse nuevamente a la distribución estadística del corpus. Algunas palabras aparecen repetidamente en varias oraciones lo cual facilita el aprendizaje por parte del alineador para decidir el emparejamiento de las palabras. Sin embargo, otras aparecen escasamente o en contextos variados lo cual hace que el alineamiento correcto sea menos probable. Nuevamente, aumentar el corpus podría prevenir o paliar estos errores. También, se ha observado la posibilidad de que la naturaleza polisintética del idioma guaraní prevenga que ciertas palabras se alineen correctamente, en consecuencia, las traducciones resultantes son de menor calidad. Por lo mismo, en un futuro, se pretende realizar un análisis morfológico de las palabras en guaraní antes de entrenar el traductor y así comprobar que al separar la raíz de las palabras de sus afijos se puede mejorar la calidad de las traducciones.

Estos fueron experimentos iniciales para tener una noción del comportamiento del Traductor Automático con el corpus recolectado. No se han hecho experimentos con otros escenarios o aplicando otras técnicas que posiblemente mejoren las traducciones actuales. Se pretende dejar estos resultados iniciales como línea base para futuras investigaciones referentes al área de traducción español-guaraní.

Conclusión

En este trabajo se ha recolectado un corpus bilingüe para el par de idiomas español-guaraní con miras a la creación de un traductor automático. Se han presentado datos de las oraciones recopiladas y las fuentes desde donde se han obtenido. Además, se ha realizado un primer intento de construir un traductor automático utilizando la plataforma de traducción Moses. Se han presentado los resultados preliminares obtenidos con el traductor de tal forma que estos sirvan como base comparativa para futuras mejoras a realizarse. Una de las mayores dificultades para avanzar con el traductor es la falta de corpus bilingüe extenso. Con este trabajo se ha dado un importante aporte en la construcción de oraciones paralelas en español y guaraní para el avance de los experimentos.

Como trabajo futuro se puede seguir ampliando el corpus. Además, se pueden aplicar otras técnicas para la mejora del traductor. Entre las técnicas posibles podemos mencionar el análisis morfológico del guaraní, la separación de las palabras en sus raíces y afijos para manejar la naturaleza aglutinante del idioma. Se ha mencionado entre los trabajos previos la existencia de un analizador morfológico que puede utilizarse para el preprocesamiento de las oraciones y luego verificar su incidencia en las traducciones resultantes.

Por otra parte, se ha demostrado que el trabajo colaborativo entre docentes y estudiantes del idioma guaraní, puede lograr el crecimiento del corpus de manera rápida y eficiente. Se propone realizar actividades como el hackathon lingüístico, de forma regular, con el fin de producir una base de datos de oraciones traducidas y validadas por entendidos del idioma.

Finalmente, se prevé la publicación del traductor como parte de una plataforma web que permita el uso del mismo y la generación de nuevas traducciones como así también el aporte de profesionales y adeptos al Guaraní para la ampliación del corpus.

El corpus bilingüe está disponible bajo solicitud para fines académicos y de investigación. Próximamente, se planea hacer público el corpus para su uso en general.

Referencias

Apertium/apertium-grn. (2020). [Python]. Apertium. https://github.com/apertium/apertium-grn (Original work published 2018)

Gasser, M. (2006). Machine translation and the future of indigenous languages. *I Congreso Internacional de Lenguas y Literaturas Indoamericanas.*

Gasser, M. (2018). Mainumby: Un Ayudante para la Traducción Castellano-Guaraní. *CoRR, abs/1810.08603*. http://arxiv.org/abs/1810.08603

Guarani Language and the Guarani Indian Tribe (Avañe'e, Jopará, Chiriguano, Mbyá). (n.d.). Retrieved March 3, 2020, from http://www.native-languages.org/guarani.htm

Hltdi/Bitext. (n.d.). GitHub. Retrieved December 1, 2020, from https://github.com/hltdi/Bitext

Koehn, P., Hoang, H., Birch, A., Callison-Burch,

C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–180. https://www.aclweb.org/anthology/P07-2045

Maldonado, D. M., Villalba Barrientos, R., & Pinto-Roa, D. P. (2016, November 22). *Eñe'ẽ: Sistema de reconocimiento automático del habla en Guaraní*. Simposio Argentino de Inteligencia Artificial (ASAI 2016) - JAIIO 45 (Tres de Febrero, 2016). http://sedici.unlp.edu.ar/handle/10915/56979

Milagros, M. P., Abdelali, A., Cowie, J., Helmreich, S., Jin, W., Ogden, B., Rad, H., & Zacharski, R. (2006). *Guarani: A Case Study in Resource Development for Quick Ramp-Up MT*.

morfo: Análisis y generación morfológica. (n.d.). Retrieved February 10, 2021, from http://plogs.soic.indiana.edu/morfo/

Moses—Main/HomePage. (n.d.). Retrieved May 12, 2020, from http://www.statmt.org/moses/

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Rudnick, A., Skidmore, T., Samaniego, A., & Gasser, M. (2014). Guampa: A Toolkit for Collaborative Translation. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1659–1663. http://www.lrec-conf.org/proceedings/lrec2014/pdf/151_Paper.pdf

SPL :: Marandukuéra. (n.d.). Retrieved May 12, 2020, from http://www.spl.gov.py/gn/index.php/marandukuera?ccm_paging_p=7

Verón, M. A., & Marae´Y, F. Y. (2018). *Tercer Seminario Internacional sobre Traducción, Terminología y Lenguas Minorizadas Jarojera Ñane Guraní Ñe'Ẽ*. http://dspace.conacyt.gov.py/jspui/handle/123456789/42606