PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Generación de corpus paralelos para la implementación de un traductor automático estadístico entre shipibo-konibo y español

Tesis para optar por el Título de Magíster en Informática con mención en Ciencias de la Computación que presenta la licenciada

Ana Paula Galarreta Asian

Asesor: Dr. Héctor Andrés Melgar Sasieta

Co-Asesor: Mag. Félix Arturo Oncevay Marcos

San Miguel, Febrero de 2017

AGRADECIMIENTOS

En primer lugar, agradezco a mi amigo Luis Fernando Muroya por motivarme a seguir este programa de maestría. También a mi madre y a mi hermana Betty por su apoyo incondicional durante esta experiencia.

Agradezco también a mi asesor, Andrés Melgar por ayudarme a determinar con claridad los objetivos de esta tesis. Además, doy las gracias a mis colegas Marco Sobrevilla y Fernando Alva, pues sus ideas y preguntas me motivaron a buscar métodos para mejorar los resultados obtenidos.

Asimismo, me gustaría agradecer a los integrantes de GRPIAA, en especial a los profesores Cesar Beltrán, Hugo Alatrista e Iván Sipiran, por inspirarme a seguir esta línea de investigación. También a Roberto Zariquiey, el maestro shipibo Juan Agustín y a todos los lingüistas que me apoyaron con la obtención y traducción de textos.

Para finalizar, agradezco especialmente a Arturo Oncevay por permitirme ser parte de este proyecto y guiarme durante todo el proceso de investigación, muchísimas gracias por tu paciencia, ideas y tiempo.

RESUMEN

Actualmente, existe información que debe estar disponible para todos los habitantes de nuestro país, tales como textos educativos, leyes y noticias. Sin embargo, a pesar que el Perú es un país multilingüe, la mayoría de textos se encuentran redactados únicamente en español. Una de las razones por las que no se traducen estos textos a otras lenguas habladas en nuestro país es porque el proceso es costoso y requiere de mucho tiempo. Por este motivo se propone desarrollar un traductor automático basado en colecciones de textos, también llamados corpus, que utilice métodos estadísticos y pueda servir de apoyo una plataforma de software de traducción automática de texto entre el español y el shipibo-konibo.

Para implementar un método estadístico, es necesario contar con corpus paralelos en los idiomas a traducir. Esto representa un problema, pues existen muy pocos textos escritos en shipibo-konibo, y la mayoría de estos no cuenta con una traducción al español. Por este motivo es necesario construir corpus paralelos en base a dos procesos: la traducción de textos del shipibo-konibo al español (y viceversa) y la alineación semi-automática de los textos bilingües disponibles. Con los corpus paralelos obtenidos, se puede entrenar y validar un traductor automático, a fin de encontrar los parámetros que generan las mejores traducciones. Además, en base a los resultados obtenidos, se determinará la etapa en la que el traductor estadístico se integrará a la plataforma de software de traducción automática que será implementada por investigadores del Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada (GRPIAA) y el departamento de lingüística de la PUCP.

ÍNDICE

I. INTRODUCCIÓN

1. MOTIVACIÓN

Existe mucha información que debe estar disponible a todas las comunidades del Perú, tales como textos educativos, leyes, noticias y planes de gobierno. Sin embargo, la mayoría de textos escritos se encuentran redactados únicamente en español, lo que origina la existencia de una brecha entre los hablantes de español y de lenguas originarias de nuestro país. En particular, en la Amazonía peruana se habla el shipibo-konibo, una lengua que cuenta con más de 30 000 hablantes (Ministerio de Cultura) y es enseñada en casi 300 colegios públicos de nuestro país (Ministerio de Educación, 2015).

En la Pontificia Universidad Católica del Perú (PUCP), el grupo de investigación RIDEI (Red Internacional de Estudios Interculturales) busca, entre otras *cosas 'establecer y reforzar canales de información y cooperación mutua entre los pueblos indígenas, los Estados, la sociedad civil y la cooperación internacional comprometida con el desarrollo de los pueblos indígenas'* (PUCP). Esta agrupación, en sus visitas a las comunidades nativas, ha identificado que existe información que debe estar disponible a todas las comunidades en nuestro país (leyes, noticias, textos educativos.). Asimismo, investigadores de la PUCP han realizado diversos estudios sobre la familia lingüística Pano y su lengua más hablada, el shipibo-konibo (Adelaar, 2011) (Zariquiey, 2006) (Zariquiey, 2011) por lo que se tiene información lingüística sobre estas lenguas. Sin embargo, a pesar de las necesidades identificadas y del interés demostrado en las comunidades amazónicas y sus lenguas, existe muy poco texto escrito en éstas.

Uno de los motivos por los que existen pocos textos en shipibo-konibo es que el proceso de traducción tradicional es costoso. Entonces, surge la necesidad de contar con una herramienta que facilite el proceso de traducción para apoyar el acercamiento del gobierno a las comunidades amazónicas y acelere este proceso, como un traductor automático. Es importante contar con material educativo en este idioma, pues actualmente existen 299 colegios públicos en Huánuco, Loreto, Madre de Dios, Ucayali y Lima en los que se enseña en shipibo-konibo en forma oral y escrita (Ministerio de Cultura, 2015).

En este contexto nace la pregunta que motiva el presente trabajo*, ¿de qué manera se puede facilitar la generación de texto en shipibo-konibo usando textos ya existentes en español?* Afortunadamente, ya existen sistemas de traducción automática en los que una computadora aprende a traducir de un idioma a otro al examinar grandes cantidades de corpus alineados (*parallel corpora*), es decir, documentos que son traducciones prácticamente exactas uno en comparación del otro. Este enfoque de traducción se conoce como basado en corpus. Sin embargo, los sistemas de traducción automática pueden utilizar también un enfoque basado en reglas o incluso, combinar ambos para tener un enfoque híbrido.

Ya que no existen corpus alineados entre el shipibo-konibo y español, se propone generarlos, ya sea a partir de textos bilingües que se puedan alinear semi-automáticamente o mediante la traducción de textos monolingües realizada por un traductor bilingüe. Posteriormente, estos textos serán utilizados para entrenar un traductor automático basado en métodos estadísticos que permita la traducción de texto entre el shipibo-konibo y español, el cuál será integrado a una plataforma de traducción híbrida desarrollada por el Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada de la PUCP en colaboración con lingüistas de esta misma universidad.

2. OBJETIVOS, METODOLOGÍA Y RESULTADOS ESPERADOS

2.1 Objetivo general

El objetivo general de este trabajo consiste *en desarrollar un traductor automático basado en métodos estadísticos que sirva de apoyo una plataforma de software de traducción automática de texto entre el español y el shipibo-konibo.*

2.2 Objetivos específicos y resultados esperados

A continuación, se mencionan los objetivos específicos y sus correspondientes resultados esperados:

2.2.1 Recolección y alineamiento textos en shipibo-konibo y en español, de forma que puedan entrenar a un traductor automático (OE1)

Se espera contar con los siguientes resultados:

a. Repositorio de documentos digitales que contengan oraciones en shipibo-konibo y su correspondiente traducción en español.
b. Archivos MS Word y MS Excel con formato adecuado para que un traductor bilingüe experto pueda realizar la traducción de texto por frases (de español a shipibo-konibo y viceversa).
c. Repositorio de documentos en formato TXT que contengan oraciones alineadas en shipibo-konibo y su correspondiente traducción en español.
d. Software que permita limpiar, extraer y ordenar oraciones alineadas provenientes de un archivo de texto.
e. Software que permita alinear un texto por frases en base a signos de puntuación.
f. Textos alineados por frases que sirvan como entrada para un traductor automático basado en corpus.

2.2.2 Desarrollar un módulo de software para la traducción automática de texto entre el español y el shipibo-konibo (OE2)

Los resultados esperados son los siguientes:

a. Determinación de los parámetros más adecuados para procesar el corpus alineado antes de utilizarlo para entrenar el traductor.
b. Traductor automático basado en métodos estadísticos que permita la traducción de frases y palabras de shipibo-konibo a español y viceversa, el cual podrá ser integrado a una plataforma de traducción híbrida.

2.2.3 Validación de la traducción de textos del español al shipibo-konibo y viceversa utilizando métricas de evaluación para traductores automáticos (OE3)

A continuación se exponen los resultados esperados:

a. Repositorio de textos traducidos del español al shipibo-konibo y viceversa.
b. Gráficos que permitan visualizar el puntaje obtenido al realizar pruebas con los parámetros sugeridos en la etapa anterior.
c. Reporte que analice el desempeño del traductor automático basado en corpus e identifique oportunidades para optimizarlo
d. Reporte que indique la etapa en la que el módulo estadístico apoyará a la plataforma de traducción híbrida.

2.3 Métodos y procedimientos

2.3.1 Metodología

Para cumplir con los objetivos expuestos en la sección 2.2 de este capítulo, se llevarán a cabo las siguientes actividades:

2.3.1.1 Elaboración de corpus paralelos (OE1)

Se buscarán y recolectarán documentos digitales que contengan texto tanto en shipibo-konibo como en español, los cuales deben ser alineados por oraciones. Luego, se deberá obtener el texto correspondiente al documento digital, lo que puede requerir de procesamiento manual. Una vez que se tengan los archivos de texto bilingüe, se desarrollará un programa que permita realizar las siguientes tareas:

1. Corregir errores y filtrar caracteres raros en un texto
2. Buscar y extraer oraciones alineadas en shipibo-konibo y español
3. Ordenar oraciones alineadas para que puedan ser procesadas con facilidad.
4. Alinear un texto por frases utilizando los signos de puntuación que se encuentran en el texto.

Por otro lado, se creará un software que permita separar textos en shipibo-konibo o textos en español por oraciones, numerarlas y colocar la información en un archivo MS Word para que puedan ser traducidas manualmente por un traductor bilingüe. Durante esta etapa, el traductor será acompañado por un lingüista, quien supervisará el proceso.

Finalmente, se agruparán todas las oraciones pertenecientes a un mismo dominio, las cuales serán ordenadas y almacenadas en archivos de texto.

2.3.1.2 Procesamiento de textos (OE1)

Para entrenar y evaluar el traductor automático, se generarán archivos de entrenamiento (*train*), afinación (*tune*) y prueba (*test*) en los idiomas de origen y destino (en este caso, shipibo-konibo y español). Sin embargo, antes de separar las oraciones de los distintos dominios en las tres partes mencionadas, se realizará un análisis de los corpus, a fin de verificar que las oraciones alineadas sean adecuadas para entrenar al traductor. Con este propósito, se analizarán dos factores: la proporción de palabras en shipibo-konibo y español y el número de palabras por oración.

2.3.1.3 Variación del número de secuencia de palabras (OE2)

Cuando una oración se traduce automáticamente de un lenguaje de entrada a un lenguaje de salida, las palabras traducidas en el proceso son reordenadas antes de generar una oración de salida. Este reordenamiento es realizado automáticamente y genera

oraciones en base al modelo de lenguaje entrenado específicamente para el lenguaje de salida.

Ya que el modelo de lenguaje puede ser entrenado en base a números distintos de secuencias de palabras, este número se variará en el entrenamiento (de 2 a 5 palabras seguidas), de forma que se puedan obtener distintos modelos del lenguaje de salida. En base a los resultados obtenidos, se determinará el número de secuencia de palabras más adecuado para la traducción de textos entre shipibo-konibo y español en los diferentes dominios.

2.3.1.4 Variación de la longitud del corpus (OE2)

En base los resultados obtenidos al variar el *modelo de lenguaje*, se modificará la longitud de los corpus para determinar hasta qué punto los resultados mejoran al tener corpus más extensos. Finalmente, se seleccionarán los parámetros más adecuados para procesar los corpus antes de introducirlos en la plataforma de traducción.

2.3.1.5 Pruebas finales (OE3)

Utilizando el traductor automático basado en corpus implementado, se realizará la traducción de textos del español al shipibo-konibo y viceversa, se almacenarán los textos obtenidos y se registrará información sobre el proceso (como tiempo de procesamiento y posibles errores de ejecución). En esta etapa también se realizarán traducciones mediante un traductor basado en reglas, el cual servirá como *baseline*. Finalmente se elaborarán reportes que muestren los resultados del proceso de validación del traductor automático basado métodos estadísticos mediante una o más métricas desarrolladas con este propósito, así como una sugerencia sobre la etapa en la que el módulo estadístico apoyará a la plataforma de traducción híbrida en base a los resultados obtenidos.

2.3.2 Métricas de traducción automática

Para evaluar el resultado de una traducción automática, se pueden realizar dos tipos de evaluación: manual y automática. En el primer caso, un evaluador (idealmente bilingüe) revisa el resultado y decide si la traducción es correcta o no. Sin embargo, no siempre se cuenta con

evaluadores bilingües, por lo que se puede utilizar un evaluador que entienda sólo el lenguaje de salida. Para evaluar el texto traducido, por lo general se verifica que el texto sea fluido y adecuado (que el significado sea el correcto). Luego, se normalizan las evaluaciones y se utilizan estos resultados para comparar diferentes traducciones.

Por otro lado, la evaluación automática compara la traducción realizada con una de referencia mediante el conteo de número de palabras que coinciden en ambas traducciones. A partir de estos resultados y de la longitud de las frases analizadas se pueden calcular valores como la precisión, que indicarán qué tan buena es una traducción. A comparación de la evaluación manual, la automática requiere de menos recursos (tanto humanos como de tiempo), lo que representa una gran ventaja.

Existen múltiples métricas de evaluación de traducción automática como NIST (Doddington, 2002), RIBES[1] (NTT Communication Science Labs, 2014), METEOR (Banerjee & Lavie, 2005), LEPOR (Aaron & Lidia, 2012) y BLEU[2] (Papineni, 2002). De todas estas, BLEU es la métrica más popular (Koehn, 2010). La forma básica de esta métrica compara secuencias de *n* palabras (*n-grams*) de una traducción automática con secuencias del mismo número de palabras de una traducción de referencia y cuenta el número de aciertos. Estos aciertos son independientes de la posición en la que aparece la secuencia de palabras. A mayor número de aciertos, mayor **precisión**. Por ejemplo, para un *unigrama* (*1-gram*) se tendría la siguiente expresión:

$$\text{precisión}_{1-\text{gram}} = \frac{\text{palabras presentes en la traducción de referencia}}{\text{palabras totales en la traducción automática}}$$

Sin embargo, los sistemas de traducción automática pueden sobre generar palabras que están presentes en la traducción de referencia sin que esta traducción sea correcta y, aun así, obtener una alta precisión, como en el siguiente ejemplo:

- *Traducción de referencia: la gata está dentro de la caja*
- *Traducción automática: la la la la la la la*
- *Precisión = 7/7*

[1] *Rank-based Intuitive Bilingual Evaluation Score*
[2] *A Bilingual Evaluation Understudy*

Entonces, se trabaja con una **precisión modificada**. En el ejemplo antes visto, *la* aparece únicamente 2 veces en la traducción de referencia por lo que, en este caso, se divide este número entre el número de palabras totales en la traducción automática, con lo que se obtiene una precisión modificada igual a *2/7*. Esta idea se puede generalizar para secuencias de más de una palabra, con lo que se puede calcular la precisión para muchas oraciones

Además, como las traducciones automáticas que son más largas que las traducciones de referencia ya son penalizadas por la precisión modificada, no es necesario penalizarlas de nuevo. Sin embargo, sí se incluye una **penalización por brevedad** (*brevity penalty*):

$$PB = \begin{cases} 1 & si\ c > r \\ e^{(1-r/c)} & si\ c \leq r \end{cases},$$

donde c es la longitud de la traducción automática y r la longitud de la traducción de referencia.

Finalmente, BLEU se expresa de la siguiente manera:

$$BLEU = PB \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

En la versión básica de BLEU, el valor por defecto del número máximo de palabras en una secuencia es igual a 4 (es decir, N=4) y los pesos se uniformizan como $w_n = 1/N$.

II. MARCO CONCEPTUAL

En esta sección se realizará una explicación sobre qué es traducción automática y cuáles son los diferentes enfoques que se pueden utilizar al traducir texto de un lenguaje origen a un lenguaje destino. Se hará énfasis en el enfoque de traducción automática basada en corpus, ya que es el que se utilizará en este trabajo.

1. TRADUCCIÓN AUTOMÁTICA

En 1954, Peter Sheridan de IBM y Paul Garvin de la universidad de Georgetown tradujeron una muestra selecta de 49 oraciones del ruso al inglés a partir de un vocabulario de 250 palabras y sólo 6 reglas gramaticales. Esta exposición atrajo la atención de los medios de comunicación en los Estados Unidos y, a pesar de tener poca relevancia científica, impulsó la inversión en traducción automática a gran escala. Desde entonces, el interés sobre estos procesos ha ido en aumento (Hutchins, 2007).

La **traducción automática** (*Machine Translation* o *MT*) consiste en traducir un lenguaje humano a otro utilizando computadoras. Para poder realizar esta tarea, los sistemas de traducción automática necesitan la siguiente información: la computadora debe conocer sinónimos de palabras y frases, así como la gramática y semántica de los lenguajes a traducir (Al-Onaizan, 1999). Una manera de incorporar este conocimiento en una computadora es mediante el uso de expertos bilingües que traduzcan la información para que pueda ser comprendida por un programa. Sin embargo, también se puede hacer que la computadora aprenda esto automáticamente al examinar grandes cantidades de **corpus alineados** (*parallel corpora*), es decir, documentos que son traducciones prácticamente exactas uno en comparación del otro.

Los sistemas de traducción automática pueden utilizar tres enfoques: MT basada en reglas, MT basada en corpus y enfoque híbrido. Se presenta un resumen de estas categorías en la Tabla N°1.

Tabla N°1. Enfoques utilizados en traducción automática

Basada en reglas *(rule-based machine translation, RBMT)*	Basada en diccionarios (Dictionary based)	Palabra por palabra
	Basada en transferencia (Transfer based)	Análisis, transferencia y generación
	Interlingua	Análisis y generación
Basada en corpus *(data-driven)*	Basada en ejemplos (Example based machine translation, EBMT)	Traducción directa, busca patrones
	Basada en métodos estadísticos (Statistical machine translation, SMT)	Encuentra la traducción más probable
Híbrido	Guiado por reglas	Usa grandes lexicones y reglas definidas por expertos
	Guiado por corpus	Aprende automáticamente a partir de corpus paralelos

1.1. Traducción automática basada en reglas (RBMT)

Bajo este enfoque, se programan reglas definidas manualmente en base a conocimiento de expertos para poder traducir texto de un lenguaje origen a un lenguaje destino (Costa-Jussa, 2015). La traducción automática basada en reglas puede ser subdividida en tres categorías:

a. ***Basada en diccionarios*** *(dictionary-based)*. Traduce palabra por palabra, requiere una sola transformación entre el lenguaje origen y el lenguaje destino.

b. ***Basada en transferencia*** *(transfer-based)*. Se da en tres etapas: análisis, transferencia y generación. Se analiza el lenguaje origen para determinar su estructura, la cual es transferida a una estructura adecuada para poder generar texto en el lenguaje destino y finalmente se genera este texto.

c. ***Interlingua.*** Consiste en dos pasos: análisis y generación. El análisis transforma el lenguaje origen en la representación interlingua (se extrae el significado) y la generación transforma la representación interlingua en el lenguaje destino. Interlingua es una representación universal de todos los lenguajes, por lo que no necesita una etapa de transferencia.

1.2. Traducción automática basada en corpus

Utiliza información proveniente de textos alineados y algoritmos de mucha complejidad para crear un modelo de traducción. Puede estar basada en ejemplos o en métodos estadísticos.

1.2.1 Traducción automática basada en ejemplos (EBMT)

Este enfoque realiza una traducción directa por analogía y puede ser considerado como un método que resuelve problemas de emparejamiento de patrones.

1.2.2 Traducción automática basada en métodos estadísticos (SMT)

La traducción automática basada en métodos estadísticos consiste en traducir texto de un lenguaje humano a otro mediante una computadora que ha aprendido cómo realizar este proceso a partir de grandes cantidades de texto traducido (Koehn, 2010).

Los modelos probabilísticos en SMT utilizan corpus paralelos para ser entrenados. Los corpus paralelos son textos en un lenguaje que se encuentran traducidos en uno distinto. Con estos textos como entrada, los modelos probabilísticos asignan un puntaje a cada traducción posible y de esta manera se encuentra la traducción con la mejor puntuación. En una revisión del estado del arte realizada por Koehn (Koehn, 2010), se encontró que la traducción automática es mejor cuando se trabaja con lenguajes similares entre sí y cuando los corpus de entrenamiento son muy grandes y constan de millones de palabras. Sin embargo, no todos los lenguajes cuentan con corpus traducidos que sean lo suficientemente extensos, por lo que son conocidos como **lenguajes de pocos recursos** (*less-resourced languages*), como por ejemplo, el shipibo-konibo.

Cuando se cuenta con suficientes textos traducidos, es necesario alinearlos para poder entrenar al traductor automático. Esta alineación puede ser por palabras o por frases. Una alineación por palabra puede ser obtenida de manera automática, pero debe ser evaluada para determinar qué tanto se parece a la alineación que haría un humano. Sin embargo, la alineación por frases es más utilizada ya que es más conveniente tener frases como **unidades atómicas** o *tokens* que tener palabras. De esta manera, se evitan problemas como el tener que traducir una palabra en un idioma que tiene como traducción a dos palabras en un idioma diferente. Por otro lado, al utilizar este método, también se reducen los problemas de ambigüedad. A continuación, se explicarán ambos tipos de alineación:

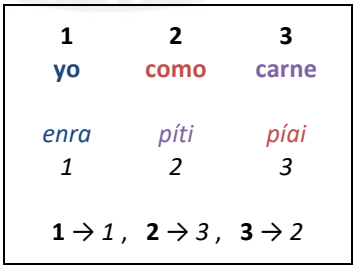
1.2.2.1 SMT basada en palabras

Los primeros trabajos en traducción automática basada en métodos estadísticos utilizaron este modelo. Si bien este enfoque ya no es parte del estado de arte, muchos de sus métodos aún son aplicados hoy en día.

Empezaremos describiendo una traducción léxica, en la que las palabras se traducen individualmente mediante el uso de un diccionario. Muchas palabras tienen más de una traducción, algunas más probables que otras. De esta forma, si se tiene un gran conjunto de corpus paralelos, se puede contar cuántas veces una palabra es traducida a otra palabra específica en un idioma. Por ejemplo, la palabra *béne* (del shipibo-konibo) puede significar *macho*, *esposo*, *marido* o *alegre* en español. Entonces, si se tienen corpus paralelos en shipibo y español, se puede contar cuántas veces la palabra *béne* es traducida a cada una de las opciones en español.

Después de hacer el conteo, se puede estimar la distribución probabilística de la traducción léxica. Así, para cada palabra en shipibo-konibo (en este caso *béne*) se devuelve una probabilidad para cada opción en español. Además, hay que tener en cuenta que cuando se traduce una oración de un lenguaje a otro, hay un **alineamiento** (*alignment*) implícito. En la Imagen N°1 se muestra un ejemplo de alineamiento en el que todas las palabras en un lenguaje tienen una correspondencia en el otro idioma.

Imagen N°1. Ejemplo de alineamiento entre español y shipibo-konibo. Se observa la alineación de las palabras *yo* con *enra*, *como* con *píai* y *carne* con *píti*.



Sin embargo, si bien generalmente una palabra en un lenguaje se traduce a una sola palabra en otro lenguaje, no siempre es así. Por ejemplo, el español tiene palabras sin un equivalente en shipibo-konibo, las cuales simplemente deben ser descartadas durante la traducción. En este caso, puede que se necesite añadir un *token* nulo (NULL) que es tratado

como cualquier otra palabra. Un ejemplo de este caso particular de alineamiento se puede observar en la Imagen N°2.

Imagen N°2. Ejemplo de alineamiento entre español y shipibo-konibo (con tokens nulos). Las palabras *dentro* y *casa* están alineadas con *meran* y *xobo*, respectivamente. Sin embargo, en este ejemplo en particular, las palabras *de* y *la* no tienen correspondientes en shipibo-konibo.

1	**2**	**3**	***4***
dentro	***de***	***la***	***casa***
	xobo	meran	
	1	2	

1 → *2* , **2** → NULL , **3** → *NULL* , *4* → *2*

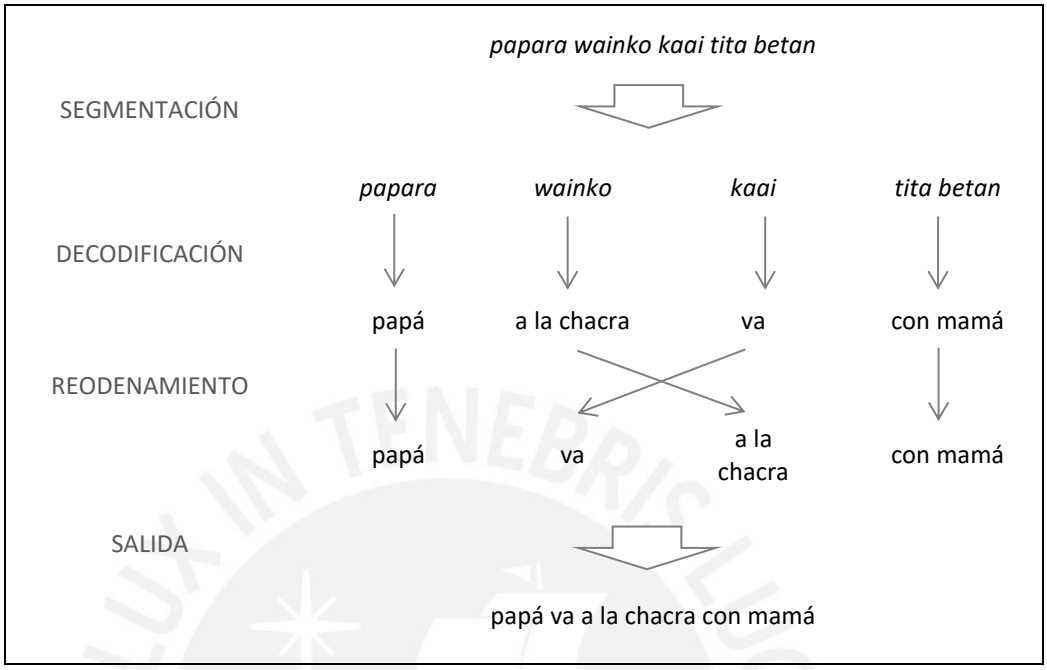
Los métodos descritos son la base de un modelo de alineamiento basado en palabras que permite descartar, añadir y duplicar palabras durante la traducción.

1.2.2.2. SMT basada en frases

Actualmente, los modelos de traducción automática más exitosos utilizan frases como unidades atómicas (Koehn, Statistical Machine Translation, 2010), las cuales están compuestas por una secuencia continua de palabras, no necesariamente por entidades lingüísticas. El procedimiento más simple a seguir es cuando se utiliza un modelo de traducción automática consiste en segmentar la oración de entrada, decodificarla y reordenar las frases para obtener la oración de salida. En la Imagen N°3 se presenta un ejemplo del proceso de traducción basado en frases.

1. **Segmentación** (*segmentation*): La oración de entrada es segmentada en frases (unidades de al menos una palabra). Todas las segmentaciones son igualmente probables y la manera en la que se segmenta la oración de entrada depende de la traducción, el reordenamiento y los puntajes del **modelo de lenguaje** (*language model*), que mide la probabilidad de que una frase sea utilizada por un hablante nativo.

Imagen N°3. Ejemplo del proceso de traducción basado en frases Primero se segmenta en frases una oración de entrada. Luego, en la etapa de decodificación, las frases son traducidas del idioma de entrada al de salida. Finalmente, se reordenan las frases traducidas y se genera una salida.



2. **Decodificación** (*decoding*). Consiste en encontrar la traducción con mejor puntaje. Por ejemplo, supongamos que en un dominio específico la palabra en shipibo-konibo *wainko* se traduce mejor como *a la chacra*. Para representar esto, se tiene una tabla que mapea no sólo palabras, sino también frases, a esta tabla se le llama tabla de traducción de frases (*phrase translation table*). Para el caso de *wainko* se podría tener la Tabla N°2.

Tabla N°2. Tabla de traducción de frases para la palabra *wainko*. En este caso, la traducción *a la chacra* tiene una probabilidad de 0.7, *de la chacra* tiene probabilidad 0.2 y *de su chacra* tiene probabilidad 0.1.

Traducción	Probabilidad
a la chacra	0.7
de la chacra	0.2
de su chacra	0.1

Es importante resaltar que los modelos actuales no se basan en el concepto de la frase, sino que agrupan de acuerdo a su aparición en el corpus. Entonces, al traducir un corpus paralelo extenso, se pueden aprender frases más largas y útiles, o inclusive, se puede memorizar la traducción de frases enteras.

3. **Reordenamiento** (*reordering*). Es dirigido principalmente por el modelo de lenguaje. Se busca que el resultado de la traducción se asemeje lo más posible a lo que diría un hablante nativo en el lenguaje objetivo pero al mismo tiempo, que este resultado se genere rápidamente. Matemáticamente, a cada oración obtenida se le asigna una probabilidad que indica qué tan factible es que ocurra en un texto.

III. ESTADO DEL ARTE

Dado que continuamente se realizan nuevos descubrimientos en el área de traducción automática, en marzo del 2016 se realizó una revisión del estado del arte para determinar cuáles de los métodos y herramientas desarrollados hasta el momento podrían ser útiles para implementar un traductor automático basado en métodos estadísticos. Como punto de partida, se revisó el libro *Statistical Machine Translation* de Philipp Cohen, quien fue uno de los pioneros en traducción automática basada en métodos estadísticos y cuyo grupo de investigación desarrolló la plataforma MOSES para traducción automática (la cual es ampliamente utilizada). A partir de la información encontrada en el libro mencionado, se buscó información sobre las herramientas típicamente utilizadas en traducción automática, cuyas funcionalidades se detallan en la primera parte de esta sección.

Por otro lado, se realizó una revisión sistemática en marzo del 2016, la cual fue actualizada en diciembre de 2016. La búsqueda se enfocó en traducción automática basada en métodos estadísticos para lenguajes de pocos recursos (como el shipibo-konibo). Se tomaron algunas partes de la revisión sistemática pero no se generó ningún protocolo. Los resultados encontrados se detallan en la segunda parte de esta sección.

1. HERRAMIENTAS DE TRADUCCIÓN AUTOMÁTICA

Algunos investigadores en traducción automática han creado herramientas que pueden ser utilizadas y modificadas por otros investigadores, ya que el código fuente típicamente está disponible (Koehn, 2016). A continuación, se describe brevemente el uso de algunas de estas herramientas.

1.1 GIZA Y GIZA++

Son herramientas que permiten alinear corpus paralelos. GIZA es parte del paquete de herramientas de traducción automática EGYPT que extrae información lingüística del corpus paralelo y fue desarrollado por el equipo *Statistical Machine Translation* durante la escuela de verano del *Center of Language and Speech Processing* en la Universidad Johns-Hopkins (Al-Onaizan, 1999). GIZA++ es una extensión del programa GIZA e incluye múltiples funciones adicionales (Och).

1.2 KenLM, IRSTLM Y SRILM

Se utilizan para elaborar los modelos de lenguaje. Al comparar los tres algoritmos (Heafield, Pouzyrevsky, Clark, & Koehn, 2013), se encontró que KenLM (Heafield K.) es más rápido que SRILM (SRI International, 2016) e IRSTLM (HLT Machine Translation, 2013).

1.3 UCAM-SMT

Es un sistema de traducción automática basado en métodos estadísticos desarrollado por la Universidad de Cambridge. Utiliza un sistema de traducción automática jerárquico basado en frases (Cambridge SMT System, 2016).

1.4 MOSES

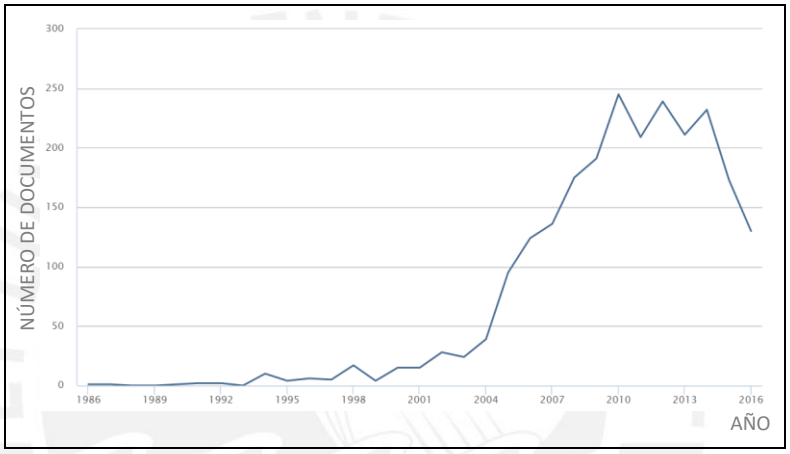
Es un sistema de traducción automática basado en métodos estadísticos que permite entrenar modelos de traducción para cualquier par de lenguajes. El modelo es entrenado a partir de una colección de corpus paralelos. Luego, un algoritmo de búsqueda eficiente encuentra la traducción más probable a partir de un número exponencial de opciones (Koehn, 2016). MOSES integra herramientas como GIZA++, KenLM y una implementación de la métrica BLEU.

2. REVISIÓN SISTEMÁTICA

Al realizar una búsqueda de artículos sobre traducción automática basada en métodos estadísticos o en corpus, se encontró que han sido publicados por lo menos 2340 artículos sobre este tema en Scopus (búsqueda realizada el 12 de diciembre del 2016), 83% de los cuales fueron publicados en los últimos diez años. La cadena de búsqueda utilizada es la siguiente:

TITLE-ABS-KEY ((("statistical" OR "corpus-based" OR "data-driven") "machine translation"))

Imagen N°4. Documentos (en SCOPUS) sobre traducción basada en métodos estadísticos vs. año de publicación. *Imagen tomada de: SCOPUS (www.scopus.com)*



Dado que se va a desarrollar un traductor para un lenguaje de pocos recursos (shipibo-konibo), se realizó una búsqueda sobre traducción automática basada en métodos estadísticos para lenguajes de pocos recursos. A continuación, se muestra la cadena de búsqueda utilizada:

TITLE-ABS-KEY ((("statistical" OR "corpus-based" OR "data-driven") "machine translation") AND (((less OR under OR low) resource) OR minority) AND language)*

Se encontraron 95 resultados en Scopus, 8 de los cuales fueron considerados como los más relevantes para el proyecto a desarrollar. A continuación, se presenta la descripción de cada uno de ellos:

En (Pérez, 2012) se detalla la generación de un corpus paralelo en español y vasco tanto para texto escrito como para lenguaje hablado, al cual denominaron EuskoParl. El texto fue obtenido a partir de reportes y discursos del Parlamento Vasco e involucra sutiles diferencias con Europarl[3]

[3] Corpus paralelos del Parlamento Europeo

(Koehn, 2005), el cual se utilizó como referencia tanto en dominio como en tamaño. Una vez que se contaba con el corpus paralelo español-vasco de 725 000 oraciones alineadas, se realizaron pruebas preliminares de traducción automática utilizando MOSES en las que se obtuvo un puntaje BLEU de 11.8% para la traducción español-vasco y 12.0% para vasco-español.

En (Mayor, 2011) se explica el desarrollo de una plataforma de traducción automática híbrida que transforma un texto del español al vasco, a la cual denominaron Matxin. Este traductor está basado principalmente en reglas y divide el proceso en tres etapas: análisis, transferencia y generación. En la primera fase, se obtiene una representación abstracta del texto en español a partir de herramientas ya existentes para procesar este idioma, las cuales permiten extraer información morfológica de cada palabra y determinar las relaciones de dependencia entre éstas. Luego, en la etapa de transferencia, se convierte la representación abstracta del texto en español (lenguaje de entrada) a una representación de texto en vasco (lenguaje de salida). Esta transferencia sucede en dos niveles: léxico (que utiliza información de diccionarios) y estructural (que usa funciones sintácticas y dependencias de frases ya identificadas). Finalmente, se genera el texto en el lenguaje objetivo. Este último paso se realiza en dos niveles: sintáctico (ordena las palabras en la secuencia correcta) y morfológico (identifica las formas de las palabras). De esta manera, se implementó el sistema Matxin, un sistema de código abierto de uso público, el cual obtuvo un puntaje BLEU de 6.3% al ser entrenado con 50 000 oraciones alineadas en vasco y español, las cuales fueron extraídas de la revista *Consumer*.

En (Pinnis & Skadiņš, 2012) se presentan técnicas para adaptar el dominio de un traductor automático estadístico mediante la utilización de términos bilingües y corpus alineados recolectados de la web tanto en inglés como en letón. Los resultados del experimento mostraron que la integración de terminología en sistemas de traducción automática estadística pueden lograr que la calidad del sistema mejore en hasta 23.1%. Por otro lado, la transformación de tablas de frases del modelo de traducción en tablas de términos que consideran el dominio utilizado puede mejorar la calidad hasta en 24.1% principalmente porque se filtran las traducciones de palabras incorrectas en el proceso de traducción. Los experimentos también muestran dos situaciones en las que no se mejora la calidad del traductor automático estadístico: cuando se usan pares de oraciones pseudo-paralelas extraídas de textos no alineados de dominios pequeños, y cuando se utilizan pares de palabras obtenidos de bancos de palabras sin realizar una desambiguación y análisis semántico del texto de entrada. Finalmente, se realizaron pruebas en un corpus extenso (5 363 063 oraciones

alineadas) y se obtuvo un puntaje BLEU de 18.21% al aplicar las técnicas mencionadas, mientras que con el baseline se obtuvo 15.85%,

Posteriormente, en (Pinnis, Skadiņa, & Vasiļjevs, 2013), los autores muestran cómo se pueden adaptar sistemas de traducción automática basados en métodos estadísticos de un dominio particular y utilizarlos con lenguajes de pocos recursos. Se realizaron experimentos para adaptar un sistema de traducción automática inglés-letón entrenado mediante corpus paralelos de dominio. Este sistema se adaptó al dominio de tecnologías de la información, añadiendo datos correspondientes a esta área que fueron extraídos de textos alineados. El sistema fue evaluado tanto por expertos como en un escenario de la vida real. En este último se mostró que la aplicación de corpus comparables generan una mejora significativa al incrementar la productividad de la traducción humana en un 13.6%, manteniendo una calidad de traducción aceptable.

En (Salami, Shamsfard, & Khadivi, 2016) se propone un modelo probabilístico para realizar traducción automática basada en métodos estadísticos. En este, utilizan un set de filtros que restringen la extracción de reglas jerárquicas a partir de parejas de frases que se pueden descomponer en dos sub-frases alineadas. Los filtros propuestos no descartan las reglas extraídas, sino que cambian el método de extracción para prevenir la selección de demasiadas reglas. De esta manera se disminuyen el tamaño del modelo, la cantidad de recursos de entrenamiento requeridos y el tiempo de decodificación. Este proceso no impacta negativamente la calidad de la traducción, sino que mejora notablemente el desempeño de modelos jerárquicos basados en frases al realizar traducciones de persa, francés y español al inglés (usaron BLEU como métrica). En el caso del traductor persa-inglés, se utilizó un corpus de aproximadamente 100 000 oraciones y los resultados mejoraron de 11.75% a 12.29%. Esta técnica puede ser aplicada para otros lenguajes, incluso para aquellos de pocos recursos para los cuales no se han desarrollado herramientas lingüísticas.

En (Chen, 2014) se utiliza un método que incorpora tanto información sintáctica del lenguaje de entrada como información morfológica del lenguaje de salida para reducir significativamente las diferencias de orden de palabras y morfología. Primero, en base al alineamiento de palabras y a los árboles sintácticos del lenguaje de entrada, se extraen automáticamente reglas de reordenamiento para obtener el orden de las palabras de la salida. Luego, en base a un modelo de Markov estadístico se adopta un método de segmentación para obtener información morfológica del lenguaje objetivo. En este estudio, se realizó una traducción entre el chino y el mongol (un lenguaje de pocos recursos) y se obtuvo como resultado un puntaje BLEU 2.1 puntos mayor que el de otros traductores

automáticos basados en frases (el puntaje más alto obtenido es 24.84%, al entrenar el traductor con 67 288 pares de oraciones chino-mongol).

En (Skadiņa, 2012) se presenta un análisis lingüístico del texto de salida obtenido en un traductor automático estadístico español-eslavo. Los autores buscaban entender completamente las principales dificultades de los sistemas de traducción automática del estado del arte, así como clasificar los principales tipos de error y analizar los motivos detrás de estos. Después del estudio, llegaron a la conclusión de que el sistema de traducción automática español-eslavo puede alcanzar una buena calidad (en el 71,5% de casos la traducción es comprensible y puede ser utilizada con pequeñas modificaciones) si las herramientas del estado del arte se aplican para crear un sistema que trabaje sobre un dominio que cuente con suficientes textos paralelos. Sin embargo, aún ocurren muchos errores de inflexión (en 57% de traducciones) y si bien formas incorrectas de palabras generan los errores más frecuentes en el corpus de entrenamiento, el impacto de este tipo de error varía entre casos y es basado en sintaxis, por lo que no se puede resolver a través de mejoras en el modelo de lenguaje o por factores morfológicos. Finalmente, encontraron dos formas en las que se puede mejorar la salida de un sistema de traducción automática estadística: aplicando técnicas basadas en sintaxis e incorporando información lingüística adicional a través de un enfoque híbrido.

En (Pa, 2016) se aplican diversos métodos de traducción automática del estado del arte para la traducción entre inglés y lenguajes de bajos recursos como lao, myanmar y tailandés (en ambas direcciones). El desempeño de los sistemas de traducción automática fue evaluado mediante las métricas BLEU y RIBES. En este estudio se observó que, en general, los sistemas SMT basados en frases tienen puntajes BLEU más altos, y los autores creen que esto indica que este método es más robusto a las limitaciones dadas por el tamaño del corpus. Sin embargo, cuando se utilizó la métrica RIBES, los mejores resultados provinieron de otros métodos, lo cual indica que estos son mejores para manejar el reordenamiento de palabras incluso cuando se tiene una cantidad limitada de datos. Finalmente, concluyen provisionalmente que los sistemas SMT basados en frases parecen ser más robustos para entrenar cantidades limitadas de datos, pero aún tienen problemas con el orden de las palabras.

En suma, se han implementado y optimizado múltiples sistemas de traducción automática para lenguajes de bajos recursos en los que se incorporan tanto métodos estadísticos como reglas lingüísticas con el objetivo de brindar el mejor resultado posible de acuerdo a métricas de evaluación (principalmente BLEU). En base a las referencias citadas, se encontró que el corpus más

reducido consta de 50 000 oraciones alineadas y tiene puntaje BLEU de 6.3% (Mayor, 2011), mientras que el más extenso consta de más de 5 millones de palabras con un puntaje BLEU mucho más alto (24.1%). Ya que los traductores entrenados con corpus más extensos tienden a obtener puntajes BLEU más altos, se hace evidente la necesidad de generar un corpus lo más extenso posible antes de empezar a entrenar un traductor basado en métodos estadísticos.

IV. ACTIVIDADES REALIZADAS

Tal como se explicó en la sección 2.3.1 del primer capítulo, se llevarán a cabo las siguientes actividades, cuya secuencia se puede observar también en la Imagen N°5:

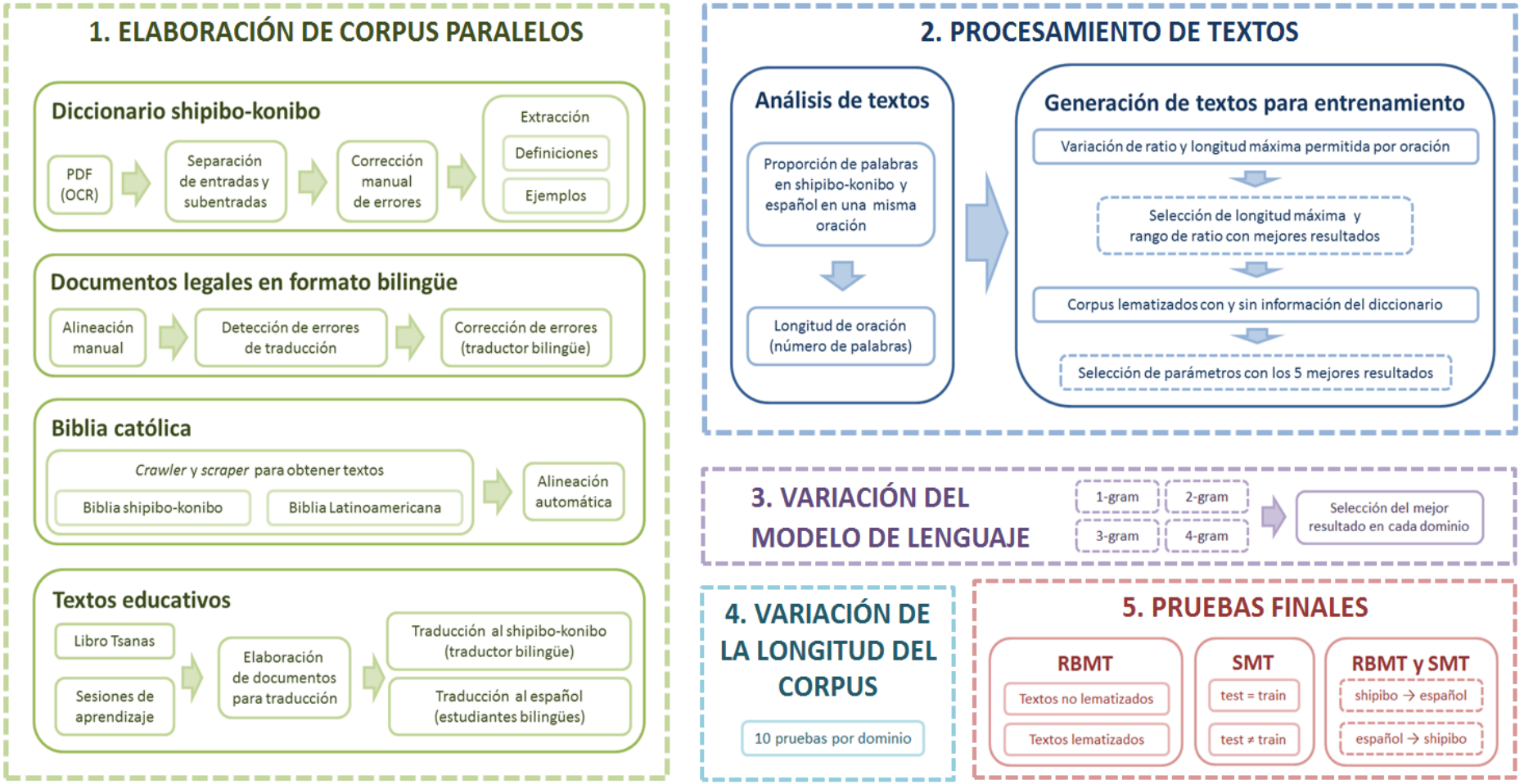
1. Elaboración de corpus paralelos
2. Procesamiento de textos
3. Variación del número de secuencia de palabras (o variación del *modelo de lenguaje*)
4. Variación de la longitud del corpus
5. Pruebas finales

Estas actividades pueden ser agrupadas en dos categorías: elaboración de corpus paralelos y experimentos realizados.

1. ELABORACIÓN DE CORPUS PARALELOS

Para poder implementar un traductor basado en métodos estadísticos se requiere de corpus paralelos. Estos textos deben estar bien alineados y sin errores de escritura (textos limpios) ya que a pesar de que los modelos de traducción automática suelen ser robustos ante datos con ruido (como aquellos textos con errores de alineación), se ha mostrado que al limpiar estos datos, los resultados mejoran (Vogel, 2003). Además, cuando se tiene poco texto disponible, es aún más importante limpiarlos y mejorarlos, ya sea aumentando las tablas de traducción de frases para incluir todas las palabras o recortando oraciones demasiado largas (Mermer, 2017).

Imagen N°5. Esquema de las actividades realizadas para generar corpus paralelos y desarrollar un traductor automático basado en métodos estadísticos del shipibo-konibo al español y viceversa



Al iniciar este proyecto, no se contaba con corpus que presentasen textos en shipibo-konibo y español en paralelo, por lo que fue necesario recolectar, digitalizar, corregir y alinear una serie de libros y documentos utilizados por el pueblo shipibo-konibo.

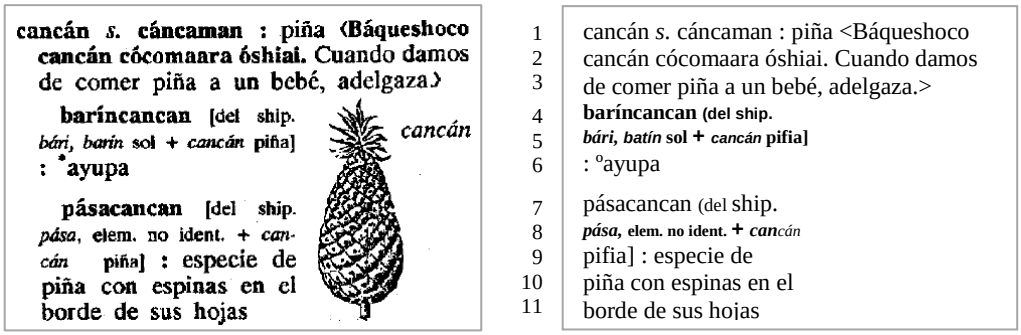
Se recolectaron textos digitales pertenecientes a 4 dominios: Diccionario, legal, religioso y educativo. Para procesar estos textos, se elaboraron múltiples programas (*notebooks)* en *Wolfram Mathematica 11* los cuales permiten limpiar, extraer y ordenar oraciones alineadas provenientes de un archivo de texto. Asimismo, se crearon *notebooks* adicionales para alinear estos textos automáticamente por frases en base a signos de puntuación.

A continuación, se explicará detalladamente el procedimiento llevado a cabo para elaborar corpus paralelos en los 4 dominios mencionados.

1.1 Diccionario Shipibo-Castellano

El primer corpus que fue elaborado consiste en ejemplos tomados del *Diccionario Shipibo-Castellano* editado por Mary Ruth Wise (Wise, 1993). Sin embargo, sólo se contaba con una versión escaneada del documento a la que se le aplicó un OCR, por lo que había un gran número de errores en el texto, tal como se muestra en la Imagen N°6. Ya que los errores sólo podían ser corregidos manualmente, se dividió el documento en cien archivos de texto, los cuales fueron corregidos voluntariamente por estudiantes de lingüística e ingeniería informática.

Imagen N°6. A la izquierda: Entrada del diccionario. A la derecha: entrada sin corregir. Se pueden ver los errores en las líneas 4 (paréntesis en vez de corchetes), 6 (° en vez de *) y 9 (pifia en vez de piña).



1 **cancán** *s.* **cáncaman : piña ‹Báqueshoco**
2 **cancán cócomaara óshiai.** Cuando damos
3 de comer piña a un bebé, adelgaza.›
4 **baríncancan** [del ship.
5 *bári, barín* sol + *cancán* piña]
6 : *ayupa
7 **pásacancan** [del ship.
8 *pása*, elem. no ident. + *can-*
9 *cán* piña] : especie de
10 piña con espinas en el
11 borde de sus hojas

cancán

1 cancán *s.* cáncaman : piña <Báqueshoco
2 cancán cócomaara óshiai. Cuando damos
3 de comer piña a un bebé, adelgaza.>
4 **baríncancan (del ship.**
5 ***bári, batín* sol + *cancán* pifia]**
6 : °ayupa
7 pásacancan (del ship.
8 ***pása,* elem. no ident. + *can***cán
9 pifia] : especie de
10 piña con espinas en el
11 borde de sus hojas

Luego, estos documentos fueron unificados en un solo archivo de texto y, utilizando patrones de texto, fueron procesados automáticamente para extraer información relevante del diccionario, basándose en la estructura de cada entrada, la cual se detalla en la Tabla N° 3.

Finalmente, se obtuvo un archivo de formato CSV con 6 252 entradas tabuladas, 3 514 de las cuales tienen al menos una ilustración verbal. A partir de este documento, se extrajeron las ilustraciones verbales (5 143 en total) y definiciones para elaborar el corpus paralelo. Como se ve en la Tabla N°4, los ejemplos se encuentran escritos entre paréntesis angulares (<>*)* y las frases en español y shipibo-konibo se separan mediante puntos. Sin embargo, algunas de las ilustraciones verbales (o ejemplos) están formadas por más de una oración, por lo que los puntos que separan oraciones en un mismo lenguaje fueron reemplazados por punto y coma para simplificar el alineamiento del corpus.

Tabla N°3. Estructura de las entradas del diccionario.

	COMPONENTE	COMENTARIO
1	Término introductorio	Palabra en shipibo (ej: *ábion*), prefijo (ej: *ra-*) o sufijo (ej: *-que*)
2	Variante del término introductorio	Categoría de la variante + variante
3	Clase	Clase de la palabra. Ejemplos: sustantivo (*s.*), pronombre (*pron.*), adjetivo (*adj.*),…
4	Parte principal	Segunda forma o parte principal del término introductorio
5	Variante de la parte principal	Categoría de la variante + variante
6	Etimología	Etimología de la palabra en corchetes. Ejemplo: *[del ship. ….]*
7	Definición	Definición o sentido equivalente en castellano. Puede ser una sola definición (*: definición*) o más de una (*1: definición_ 2: definición_2 …*)
8	Nota de uso	Nota o comentario acerca de la definición. Presenta la siguiente estructura: *-Úsase …*
9	Ilustración verbal de la definición	Palabra, frase u oración que se presenta como ejemplo. Tiene la siguiente estructura: *<shipibo . español>*
10	Subentrada	Tiene la misma estructura que la parte principal
11	Referencias	Tienen la siguiente estructura: *Véase…*
12	Párrafo de sinónimos	Sinónimos en mayúsculas e información adicional. Tiene la siguiente estructura: *sinón. …*

Tabla N°4. Ejemplo de información extraída del diccionario. No se incluyen todos los componentes de las entradas.

TERMINO	CLASE	DEFINICIÓN	ILUSTRACIÓN VERBAL
cancán	s.	piña	<Báqueshoco **cancán** cócomaara óshiai. Cuando damos de comer piña a un bebé, adelgaza.>
nemín	adj.	profundo	<Nato páro riqui, jénetian quiquínbires **nemín**; icáxbiri iqui, báritianbiribi benéshoco. El Ucayali es muy hondo en el *invierno, pero en el *verano, por el contrario, tiene muy poca agua.> <Manánquini **nemín**hainoax yoshín jóxonronqui jóni píti piánique. Cuentan que antiguamente un demonio vino de una cueva profunda y comió la comida de un hombre.>

En la Tabla N°5 se muestra un resumen del conteo de entradas del diccionario, así como las palabras y oraciones alineadas.

Tabla N°5. Conteo de entradas y oraciones en el diccionario. Se detallan el número de ejemplos, las ilustraciones verbales y palabras alineadas a partir del documento.

	SHIPIBO-KONIBO	ESPAÑOL
Entradas	6 252	
Entradas con ilustraciones verbales	3 514	
Palabras alineadas	29 625	51 399
Ilustraciones verbales alineadas (oraciones alineadas)	5 143	

1.2 Documentos legales en formato bilingüe

Se obtuvieron dos leyes y un proyecto del Estado peruano en formato PDF:

a. *Ley Forestal y de Fauna Silvestre N° 29763* (Ministerio de Agricultura) (Ministerio de Agricultura y Riego)
b. *Ley de Lenguas Indígenas u Originarias N° 29735* (ley que regula el uso, preservación, desarrollo, recuperación, fomento y difusión de las lenguas originarias del Perú) (Ministerio de Cultura, 2014)
c. *Hoja de ruta o resumen del plan de consulta previa del proyecto Hidrovía Amazónica* (Hoja de ruta o resumen del plan de consulta previa del proyecto Hidrovía Amazónica) (Teenoxon jaskakin noa yokakanti shinanbo nato Idrovia Amazonica ikainko)

Los tres documentos se encontraban redactados tanto en shipibo-konibo como en español, pero tuvieron que ser alineados manualmente debido a ciertas incongruencias en los textos (como diferente número de artículos en una misma ley). Además, se encontraron párrafos cuyo número de palabras era considerablemente mayor en español que en shipibo-konibo, por lo que los textos fueron revisados por un traductor bilingüe, quien los corrigió. Sin embargo, dado que muchos términos del dominio legal no tienen traducción directa del español al shipibo-konibo (como "ley"), y esta tarea requería de mucho tiempo, se tomó la decisión de no traducir todas las oraciones faltantes, por lo que algunos artículos de la Ley Forestal no fueron considerados. En Tabla N°6 se muestra un conteo de las palabras y oraciones alineadas en los tres documentos legales mencionados.

Tabla N°6. Conteo de palabras y oraciones en documentos legales. Se detallan las palabras en el texto digitalizado, palabras alineadas y oraciones alineadas en los tres documentos procesados.

	PALABRAS EN EL TEXTO DIGITALIZADO		PALABRAS ALINEADAS		ORACIONES ALINEADAS
	SHIPIBO-KONIBO	ESPAÑOL	SHIPIBO-KONIBO	ESPAÑOL	
Ley Forestal y de Fauna Silvestre N° 29763	19 805	24 851	17 747	22 095	853
Ley de Lenguas Indígenas u Originarias N° 29735	2 049	2 919	2 049	2 919	131
Proyecto Hidrovía Amazónica	1 887	2 036	1 887	2 036	158
			21 683	**27 050**	**1 142**

1.3 La biblia católica

Se empleó una traducción de la biblia católica a shipibo-konibo. Inicialmente, se trabajó con la versión en shipibo-konibo perteneciente al sitio web *Digital Bible Society* (Wycliffe Bible Translators, Inc, 2012), que ofrece traducciones aproximadas de la biblia en diferentes lenguas. Sin embargo, la extracción de los textos de forma automática resultó ser muy complicada, por lo que finalmente se decidió utilizar la versión de *Word Bibles* (World Bibles). Por otro lado, la biblia en español corresponde a la *Biblia Latinoamericana,* extraída de la página web de la *librería San Pablo* (San Pablo España).

Para obtener los textos, se empleó un *crawler,* pues cada capítulo de la biblia se encontraba en una dirección URL diferente. En total, se obtuvieron 328 libros en shipibo-konibo y 1 334 en español, a partir de los cuales se alinearon automáticamente 9 804 versículos, pues no todos los capítulos en español tenían traducción al shipibo-konibo. Finalmente, se dividió el texto de los versículos utilizando signos de puntuación específicos (punto, dos puntos, signos de interrogación y signos de admiración), con lo que se obtuvieron 13 257 oraciones alineadas. Cabe resaltar que sólo se dividieron aquellos versículos que contenían el mismo número de signos de puntuación específicos tanto en su versión en shipibo-konibo como en español. Los resultados se pueden observar en la Tabla N°7.

Tabla N°7. Conteo de palabras y oraciones en la biblia católica. Se indican el número de versículos, palabras y oraciones alineadas.

	SHIPIBO-KONIBO	**ESPAÑOL**
Libros obtenidos	328	1 334
Versículos alineados (pueden contener más de una oración)	9 804	
Palabras alineadas	210 649	206 909
Oraciones alineadas	**13 257**	

1.4 Textos educativos

Los textos del dominio educativo son relevantes para este proyecto debido a que el Ministerio de Educación promueve la educación intercultural, bilingüe y rural (Ministerio de Educación, 2012). En este contexto, el gobierno peruano se ve en la necesidad de contar con textos educativos en lenguas originarias como el shipibo-konibo, por lo que se realizó un esfuerzo especial para generar corpus paralelos en este dominio.

La Dirección General de Educación Intercultural, Bilingüe y Rural ha elaborado textos en shipibo-konibo para educación inicial y primaria, los cuales están siendo utilizados para elaborar corpus paralelos en el dominio educativo. Hasta la fecha, se ha trabajado con un material de enseñanza para niños de nivel escolar primaria: *Axeti kirika Tsanas: 4 Baritiayabaona* (Libro de aprendizaje Tsanas: para niños de 4to grado de primaria) (Ministerio de Educación, 2014). Este texto se encontraba únicamente en shipibo-konibo, por lo que debía ser traducido al español. La tarea de traducción fue llevada a cabo por el maestro shipibo Juan Agustín, quien trabajó con 15 documentos

en MS Word, de hasta 100 oraciones cada uno. Dichas oraciones se mostraban en una tabla para poder simplificar la traducción y posterior recopilación de los textos alineados, tal como se puede ver en el ANEXO N°1. Finalmente, se obtuvieron 1 435 oraciones alineadas.

Por otro lado, se inició la traducción de las Sesiones de Aprendizaje 2016 (Ministerio de Educación, 2016). En el sitio web correspondiente a estas sesiones, se pueden descargar documentos divididos en unidades didácticas, en los que se información como situaciones significativas, aprendizajes esperados, evaluación y momentos de la sesión. Para traducir estos textos, se contó con el apoyo de 12 estudiantes de educación bilingüe, quienes hablan shipibo-konibo y español con fluidez. A cada uno de ellos se le entregó un documento en MS Excel en el que figuraba el texto correspondiente a la descripción de la unidad didáctica en español. Para evitar la traducción de oraciones largas, estos textos fueron separados por puntos, comas, signos de interrogación y signos de exclamación, de forma que la traducción fuese lo más literal posible. En la sesión de traducción piloto llevada a cabo el 26 de Noviembre del 2016, se lograron traducir 663 oraciones.

En la Tabla N°8 se muestra un resumen de los textos traducidos en el dominio educativo.

Tabla N°8. Conteo de palabras y oraciones en documentos del dominio educativo. Se detallan las palabras y oraciones traducidas en los documentos procesados.

	PALABRAS TRADUCIDAS		**ORACIONES TRADUCIDAS**
	SHIPIBO-KONIBO	**ESPAÑOL**	
Axeti kirika Tsanas: 4 Baritiayabaona	98 247	118 378	1457
Sesiones de Aprendizaje 2016	19 850	21 709	663
	118 092	**140 087**	**2 120**

2. EXPERIMENTOS REALIZADOS

2.1 Procesamiento de textos

Se decidió utilizar la plataforma MOSES para implementar el traductor automático debido a que integra todas las herramientas necesarias para realizar esta tarea: GIZA++, KenLM e incluso una implementación de la métrica de evaluación BLEU. Para entrenar y evaluar el traductor, se necesitan 6 archivos de texto: de entrenamiento (*train*), afinación (*tune*) y prueba (*test*) en los idiomas de origen y destino (shipibo-konibo y español). Ya que frecuentemente se utiliza la proporción 80-10-10 para separar datos de entrenamiento, afinación y prueba, se siguió este estándar en todos los experimentos realizados. Además, ya que palabras y frases pueden tener diferentes significados en dominios diferentes, la traducción debe adaptarse al estilo esperado en cada uno de estos (Koehn, 2010), por lo que los corpus deben estar separados por dominio.

2.1.1 Análisis de textos

Antes de separar las oraciones para entrenamiento, afinación y prueba, se realizó un análisis de los corpus, a fin de mejorar su calidad antes de utilizarlos para entrenar al traductor. Es posible que al realizar el alineamiento de manera semi-automática, algunas frases hayan sido alineadas de manera errónea. Además, el tiempo de entrenamiento del traductor se reduce al tener oraciones más cortas (Koehn, 2016) y, en la práctica, la longitud de las oraciones usualmente se limita a 8 palabras (Dimeo, 2014). Para determinar si las oraciones alineadas que forman parte de los corpus eran adecuadas para entrenar el traductor, se analizaron dos factores: la proporción de palabras en shipibo-konibo y español y el número de palabras por oración.

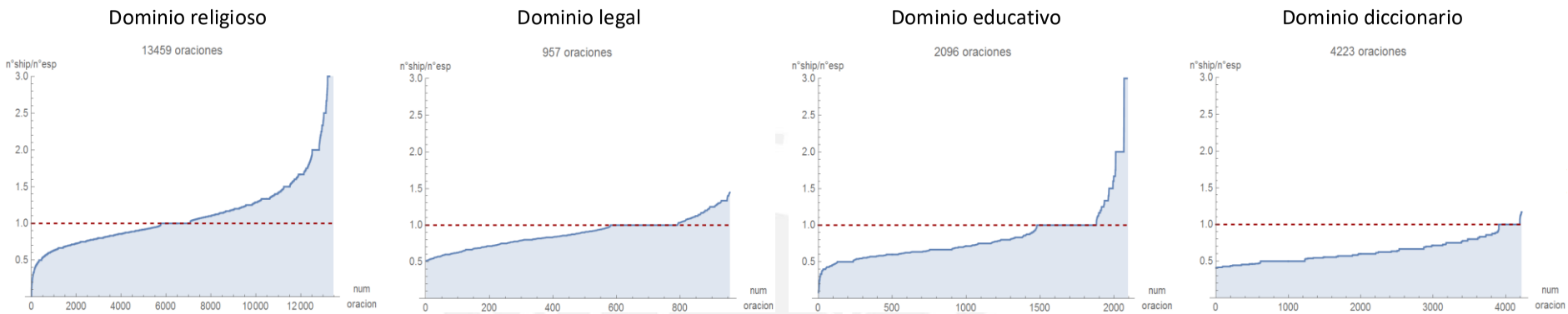
2.1.1.1 Proporción de palabras en shipibo-konibo y español en una misma oración

Se realizó un conteo de palabras en shipibo-konibo y en español en cada una de las oraciones de los corpus. Luego, se obtuvo la proporción entre estas mediante la siguiente expresión:

$$ratio_{\ oración} = \frac{n°ship}{n°esp} = \frac{número\ de\ palabras\ en\ shipibo-konibo}{número\ de\ palabras\ en\ español}$$

Se espera que en la mayoría de casos las oraciones tengan más palabras en español que en shipibo-konibo (ratio < 1) debido a que este último es predominantemente un lenguaje aglutinante (Bismark, 2016).

Imagen N°7. Proporción de palabras en shipibo-konibo y español (por oración). La línea horizontal punteada corresponde a una proporción igual a 1,0.



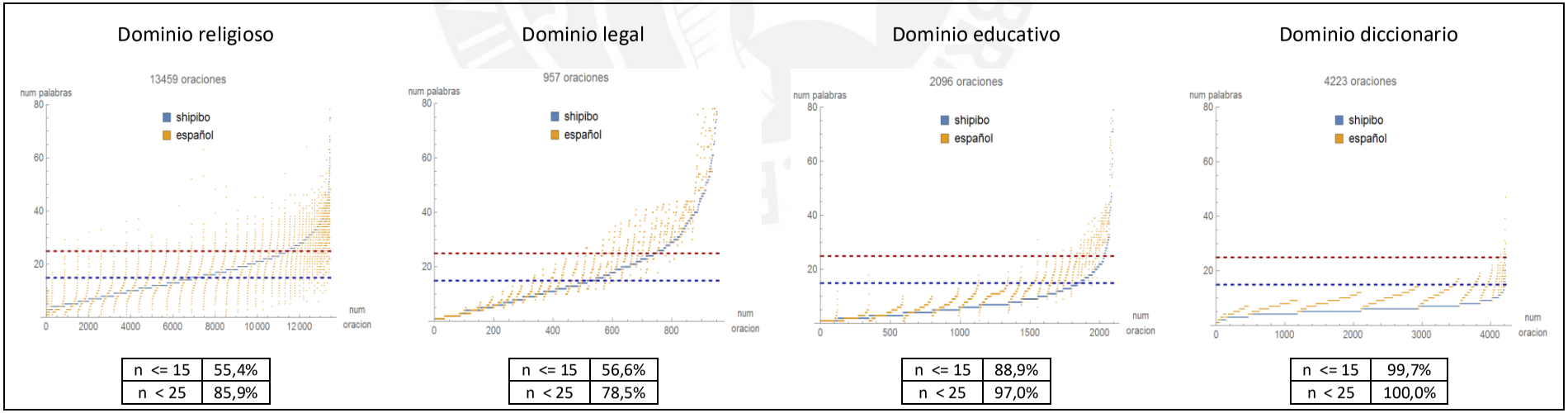
ratio > 1.0	47,4%
ratio = 1,0	9,8%
ratio < 1,0	42,8%

ratio > 1.0	17,2%
ratio = 1,0	22,2%
ratio < 1,0	60,6%

ratio > 1.0	10,3%
ratio = 1,0	19,2%
ratio < 1,0	70,5%

ratio > 1.0	0.7%
ratio = 1,0	6,8%
ratio < 1,0	92,5%

Imagen N°8. Número de palabras en shipibo-konibo y en español (por oración). La línea horizontal azul corresponde a un número de palabras en shipibo-konibo igual a 15 y la roja a 25.



n <= 15	55,4%
n < 25	85,9%

n <= 15	56,6%
n < 25	78,5%

n <= 15	88,9%
n < 25	97,0%

n <= 15	99,7%
n < 25	100,0%

Los resultados obtenidos se muestran en la Imagen N°7. Se puede observar lo siguiente sobre los cuatro dominios:

a. *Dominio religioso:* La proporción de palabras está balanceada: 47.4% tiene más palabras en shipibo-konibo y 42.8% tiene más palabras en español. El 9.8% restante tiene el mismo número de palabras. Esta distribución probablemente se deba a que los misioneros buscaron realizar una traducción literal de la biblia, por lo que el número de palabras en español y shipibo-konibo no varía mucho en una misma oración.
b. *Dominio legal:* En la mayoría de las oraciones, el número de palabras en shipibo-konibo es menor o igual al número de palabras en español y son pocas las oraciones en la que el número de palabras en shipibo-konibo es mayor (menos del 20%). Al comparar manualmente las oraciones alineadas, se observó que cuando se tradujeron los documentos legales del español al shipibo-konibo, algunas frases fueron omitidas (no se tradujeron), lo cual explicaría la proporción de palabras obtenida.
c. *Dominio educativo:* Sólo el 10.3% de las oraciones tiene más palabras en shipibo-konibo que en español. Esto puede deberse a que la traducción se realizó del shipibo-konibo al español, por lo que el traductor usó más palabras en español para mantener el sentido de la oración.
d. *Dominio diccionario:* Menos del 1% de las oraciones tiene más palabras en shipibo-konibo que en español. Probablemente, las oraciones del diccionario fueron traducidas a partir de frases en shipibo-konibo recolectadas por los lingüistas. Entonces, para poder transmitir bien el significado de una frase traducida, se utilizaron más palabras en español que en shipibo-konibo.

2.1.1.2 Número de palabras en una oración:

Se determinó el número de palabras tanto en shipibo-konibo como en español en cada una de las oraciones para determinar si es factible limitar a 8 el número de palabras por oración. El análisis de este conteo se muestra en la Imagen N°8. Al observar el número de palabras en shipibo-konibo en los diferentes dominios, se encontró que el dominio *Diccionario* tenía las oraciones más cortas, siendo la más larga una oración conformada por 22 palabras, la cual se puede ver en la Tabla N°9.

Tabla N°9. Oración más larga en shipibo-konibo en el dominio diccionario. Se muestra también su traducción al español.

ORACIÓN EN SHIPIBO-KONIBO	ORACIÓN EN ESPAÑOL
rínsira jáke tibítoain nóriti amirikain jaínoax orópain ja riki míxo jisá jaboen jína máxkoxoko jawen ráni riki kextó jawen bíchi riki kopíora (22 palabras)	*el lince vive en el tibet en norte américa y en europa tiene un parecido al gato pero su cola es más corta su pelaje es valioso en el mercado* (30 palabras)

2.1.2 Generación de textos para el entrenamiento del traductor automático

El desempeño del traductor automático depende, en parte, del corpus utilizado para su entrenamiento. Para generar los textos de entrenamiento, se tuvo en cuenta lo siguiente:

a. Oraciones diferentes tienen un número diferente de palabras (**longitud** de oración)
b. La proporción de palabras en shipibo-konibo y español varía (**ratio** de oración)
c. Las palabras de las oraciones pueden ser **lematizadas** automáticamente
d. Las definiciones del **diccionario** pueden ser añadidas al corpus de entrenamiento

Es por eso que se diseñaron múltiples experimentos con la finalidad de encontrar una combinación óptima de estor parámetros. Primero, se realizaron pruebas variando el ratio y la longitud de oración máxima permitida en oraciones no lematizadas y sin considerar las definiciones del diccionario. Una vez determinados los ratios y las longitudes máximas adecuadas, se utilizaron estos parámetros para entrenar las oraciones lematizadas con y sin las definiciones del diccionario.

2.1.2.1 Variación de ratio y longitud máxima permitida

En la literatura se indica que al tener oraciones más cortas se reduce el tiempo de entrenamiento del traductor (Koehn, 2016). Además, en la práctica, la longitud de las oraciones usualmente se limita a 8 palabras (Dimeo, 2014). Sin embargo, en nuestro caso no es factible realizar el corte utilizando este límite, pues el número de oraciones de entrenamiento sería muy pequeño.

Entonces, se decidió realizar múltiples experimentos de traducción del shipibo-konibo al español, variando el ratio de oración permitido para cada uno de los cuatro dominios, así como el número máximo de palabras permitidas en cada oración en shipibo-konibo. Hay que considerar que al variar estos parámetros, el tamaño del corpus también cambia.

Antes de generar los archivos de texto, se revisaron manualmente las oraciones de los valores extremos, con lo que se revelaron algunos errores en los dominios educativo y religioso:

a. *Dominio educativo.* Errores de traducción cometidos por los estudiantes de educación bilingüe. Para trabajos futuros, estos errores podrían ser corregidos manualmente por el maestro bilingüe Juan Agustín.
b. *Dominio religioso:* Errores en el alineamiento automático cuando el texto en shipibo de un versículo específico no coincide con el texto del mismo versículo en español. Probablemente, se debe a que existen versiones diferentes de la biblia. Además, es posible que algunos versículos no hayan sido alineados correctamente debido a errores tipográficos en el texto original, el cual no pudo ser corregido manualmente debido a su extensión (más de 13 000 oraciones).

Ya que los errores encontrados en los textos del dominio educativo fueron cometidos por los estudiantes de educación bilingüe, se decidió separar las oraciones traducidas por el maestro Juan Agustín y realizar pruebas sobre esta selección como si se tratase de otro dominio. Al hacer esta separación, se cuenta finalmente con cinco dominios en total.

Para seleccionar los ratios de oración con los que se realizaron las pruebas, se estableció una longitud mínima de 200 oraciones en el corpus de entrenamiento, debido a que la elaboración del modelo de lenguaje (*language model*) presenta errores con un número menor. Es decir, si se utilizan menos de 200 palabras, no se genera ningún modelo de lenguaje, sólo se obtiene un archivo de texto que indica que hubo un error durante la ejecución.

Después de eliminar oraciones duplicadas, se generaron múltiples sets de textos con diferentes ratios de oración permitidos, específicos para cada dominio. Los rangos de ratio permitidos varían en pasos de 0.1 y sus límites se pueden observar en la Tabla N°10.

Por otro lado, la longitud máxima aumentó de 5 en 5, a partir de una longitud de 10 palabras. Sin embargo, ya que la longitud máxima permitida típicamente se establece en 8 (como se mencionó anteriormente), también se realizaron pruebas con este valor. El límite superior de longitudes máximas con las que se experimentó se estableció empíricamente de acuerdo a cada dominio, con el objetivo de encontrar una longitud óptima.

Tabla N°10. Se muestra el número de oraciones obtenido al establecer diferentes límites de ratio de permitido, así como el porcentaje que estas oraciones representan en el corpus sin filtrar.

DOMINIO	LÍMITE INFERIOR	LÍMITE SUPERIOR	PORCENTAJE	NÚMERO DE ORACIONES
Religioso	0.2	2.0	97%	12 801
Diccionario	0.4	1.2	82%	4 233
Educativo	0.2	2.0	97%	2 061
Educativo 2	0.2	2.0	98%	1 432
Legal	0.5	1.5	84%	957

En esta etapa se utilizó un modelo de lenguaje que agrupa las palabras en español de dos en dos (bigramas). Y, además de los experimentos antes descritos, se realizaron pruebas sin aplicar ningún filtro ni de longitud ni de ratio, para poder tener una base con la cual comparar los resultados, los cuales se muestran en las imágenes N°9, N°10 y N°11. La implementación de BLEU utilizada proporciona los resultados en porcentajes (1 punto equivale a 1% de precisión).

En la tabla N°11 se presenta un resumen de los resultados obtenidos en esta etapa. Se observa lo siguiente:

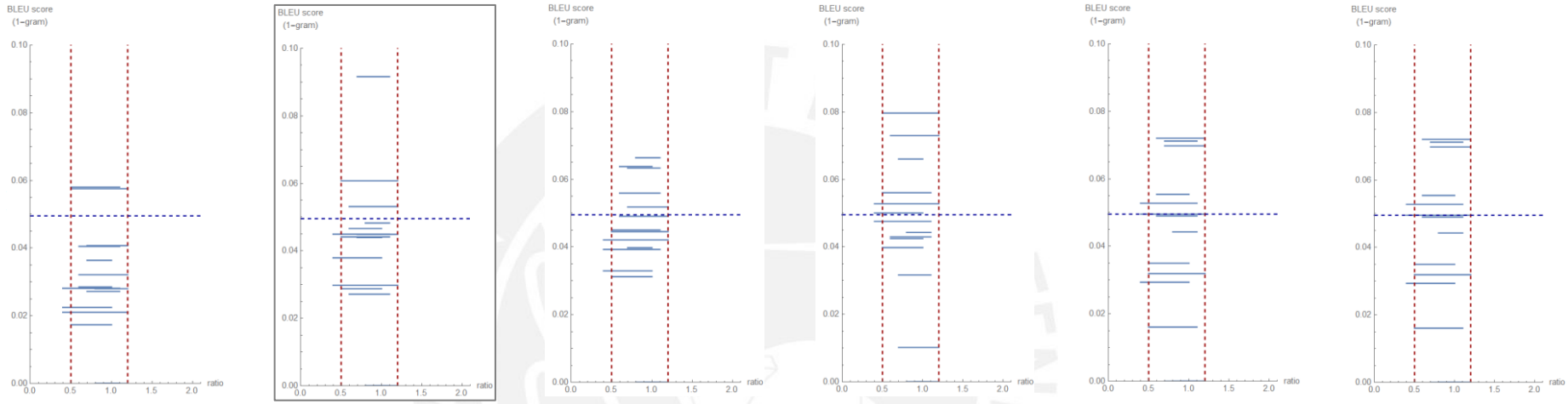
a. *Dominio Diccionario:* Los mejores resultados fueron obtenidos en aquellos ensayos con longitud máxima 10. Los ratios con mejores puntajes se encuentran entre 0.5 y 1.2.
b. *Dominio legal:* Ninguno de los ensayos con longitud máxima de oración 10 u 8 obtuvo un valor diferente de cero de acuerdo a la métrica BLEU para unigramas, es por eso que se realizaron pruebas para longitudes máximas de hasta 50 palabras. Por otro lado, ya que

los puntajes obtenidos para longitudes máximas 15 y 20 fueron bajos, no se muestran estos gráficos. Los mejores resultados se obtuvieron para una longitud máxima de 20 palabras y ratios entre 0.4 y 1.4

c. *Dominio educativo:* Se obtuvieron los mejores resultados al trabajar con el dominio educativo 2 (sólo con las oraciones traducidas por el maestro Juan Agustín) en comparación al dominio educativo 1 (todas las oraciones). En el dominio educativo 2, los mejores puntajes se dieron con longitud máxima de 20 palabras y ratios entre 0.3 y 1.5.
d. *Dominio religioso:* Los mejores resultados fueron obtenidos al establecer una longitud máxima igual a 25 y ratios entre 0.7 y 2.0. De todos los dominios analizados, el religioso cuenta con el corpus más extenso y es el que tuvo puntaje más alto (BLEU-1 igual a 0.39).

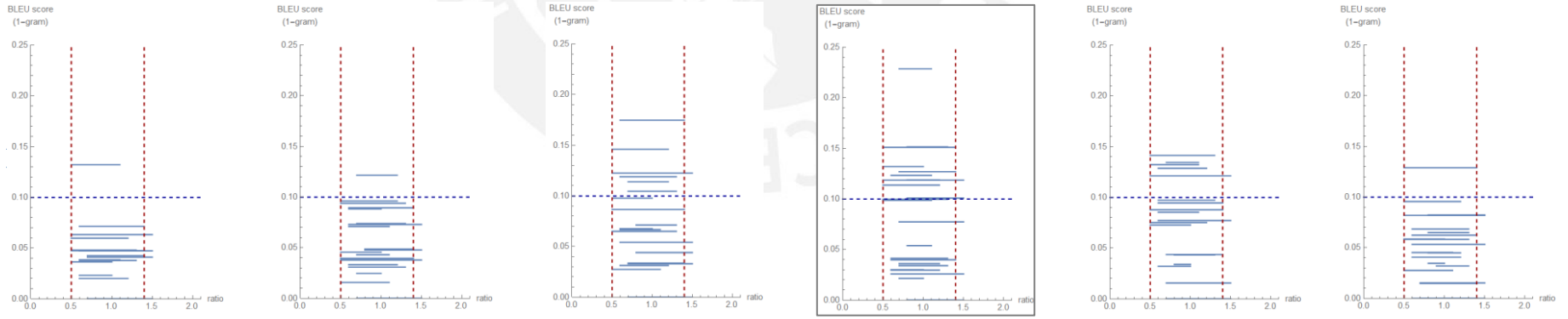
Imagen N°9. Gráficos de Puntaje BLEU (1-gram) vs. ratio de oración para el dominio Diccionario y el dominio legal. Cada línea horizontal azul claro corresponde a un rango de ratio permitido (si la línea va de 0.5 a 1.2 en el en el eje horizontal, quiere decir que se permitieron ratios entre estos valores). La línea horizontal azul punteada corresponde al resultado obtenido sin aplicar ningún filtro y las líneas rojas al rango de ratio seleccionado para realizar los siguientes experimentos.

DOMINIO DICCIONARIO



LONGITUD MÁXIMA = 8 | **LONGITUD MÁXIMA = 10** | LONGITUD MÁXIMA = 15 | LONGITUD MÁXIMA = 20 | LONGITUD MÁXIMA = 25 | LONGITUD MÁXIMA = 30

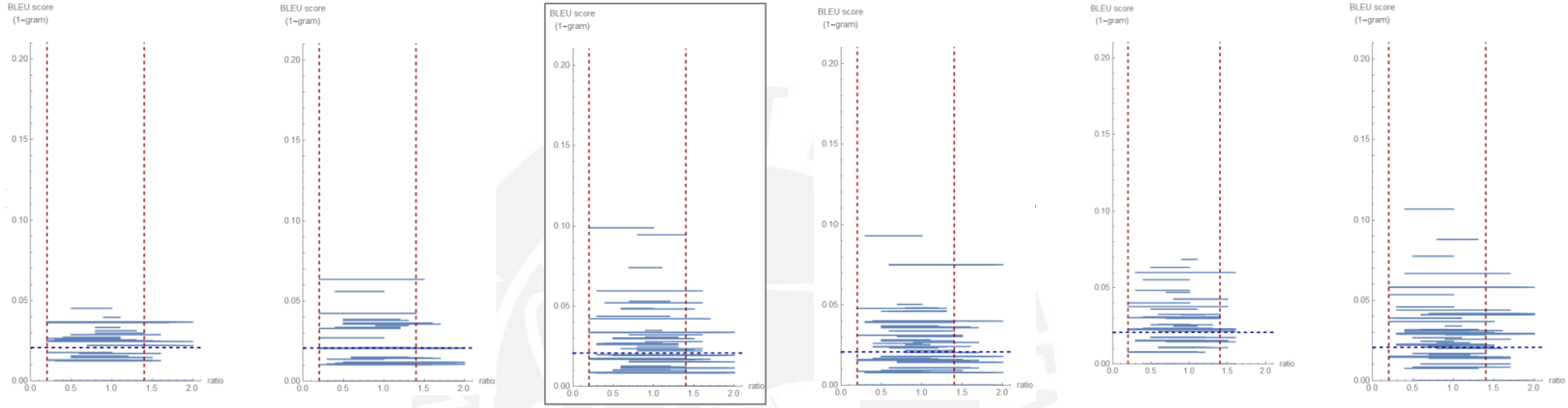
DOMINIO LEGAL



LONGITUD MÁXIMA = 25 | LONGITUD MÁXIMA = 30 | LONGITUD MÁXIMA = 35 | LONGITUD MÁXIMA = 40 | LONGITUD MÁXIMA = 45 | LONGITUD MÁXIMA = 50

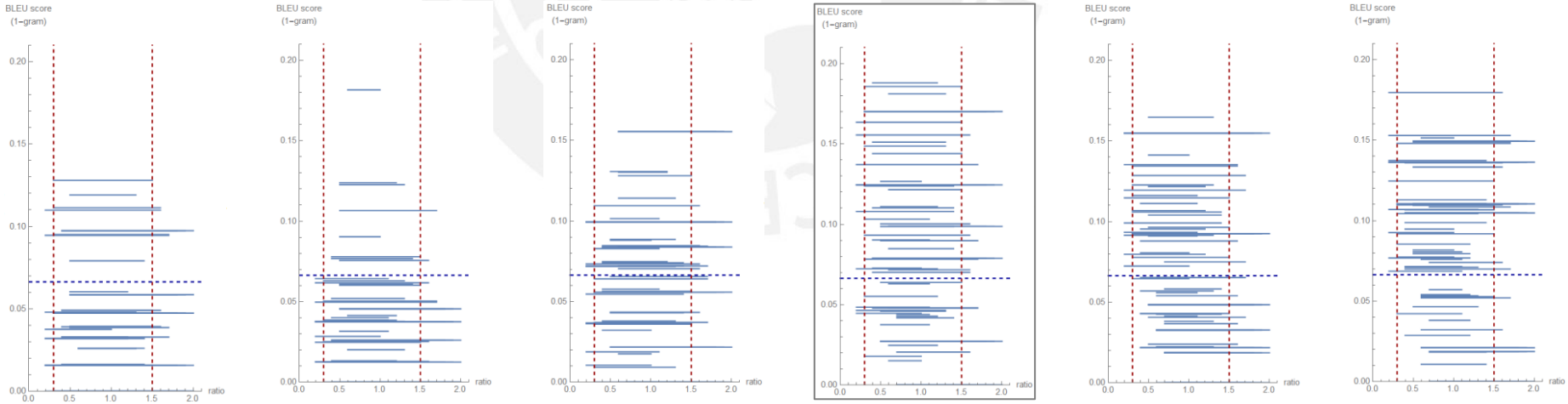
Imagen N°10. Gráficos de Puntaje BLEU (1-gram) vs. ratio de oración para diferentes límites de longitud en los dominios educativos. La línea horizontal azul punteada corresponde al resultado obtenido sin aplicar ningún filtro y las líneas rojas al rango de ratio seleccionado para realizar los siguientes experimentos.

DOMINIO EDUCATIVO 1 (TODAS LAS ORACIONES DISPONIBLES)



LONGITUD MÁXIMA = 8 | LONGITUD MÁXIMA = 10 | LONGITUD MÁXIMA = 15 | LONGITUD MÁXIMA = 20 | LONGITUD MÁXIMA = 25 | LONGITUD MÁXIMA = 30

DOMINIO EDUCATIVO 2 (ÚNICAMENTE ORACIONES TRADUCIDAS POR EL MAESTRO BILINGÜE JUAN AGUSTÍN)



LONGITUD MÁXIMA = 8 | LONGITUD MÁXIMA = 10 | LONGITUD MÁXIMA = 15 | **LONGITUD MÁXIMA = 20** | LONGITUD MÁXIMA = 25 | LONGITUD MÁXIMA = 30

Imagen N°11. Gráficos de Puntaje BLEU (1-gram) vs. ratio de oración para diferentes límites de longitud en el dominio religioso. La línea horizontal azul punteada corresponde al resultado obtenido sin aplicar ningún filtro por longitud y las líneas rojas al rango de ratio seleccionado para realizar los siguientes experimentos.

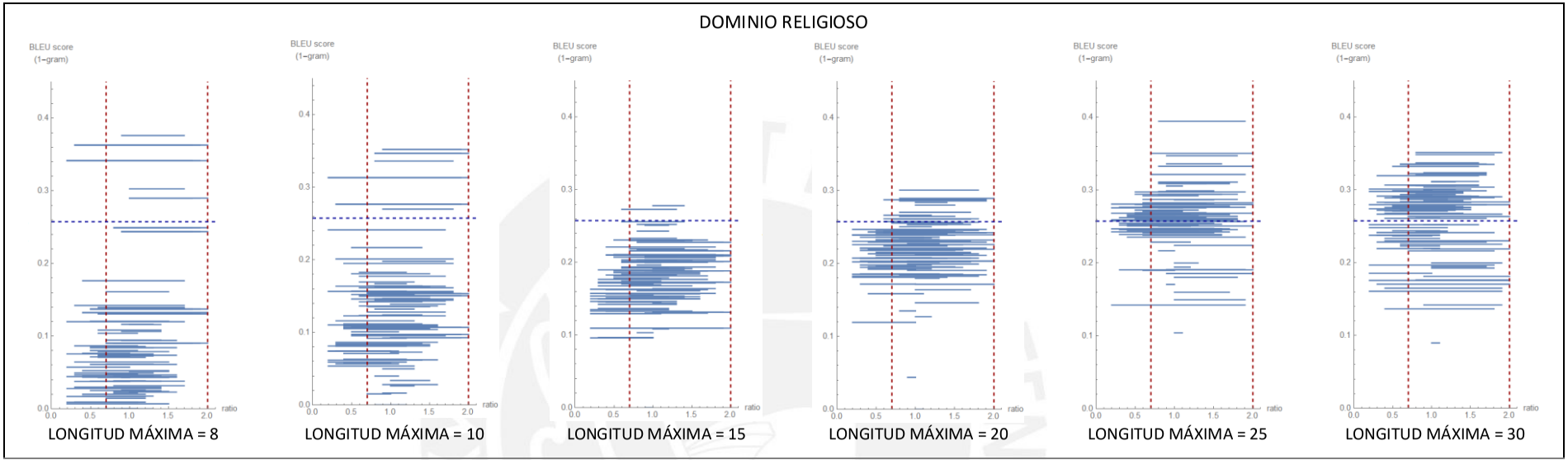


Tabla N°11. Tabla resumen de los resultados obtenidos al variar la longitud máxima permitida y el ratio en cinco dominios diferentes.

	DOMINIO DICCIONARIO	DOMINIO LEGAL	DOMINIO EDUCATIVO 1	DOMINIO EDUCATIVO 2	DOMINIO RELIGIOSO
Límite inferior en el experimento	0.4	0.5	0.2	0.2	0.2
Límite superior en el experimento	1.2	1.5	2.0	2.0	2.0
Puntaje BLEU (1-gram) más alto sin filtros	0.04953	0.09987	0.02072	0.06637	0.2575
Longitud del corpus de entrenamiento sin filtros	3 380	766	1 651	1 149	10 769
Puntaje BLEU (1-gram) más alto	0.09162	0.22866	0.09871	0.18809	0.3946
Longitud del corpus de puntaje más alto	2 259	695	1 358	1 035	7 040
Límite inferior del ratio de puntaje más alto	0.7	0.7	0.2	0.4	0.8
Límite superior del ratio de puntaje más alto	1.1	1.1	1.0	1.2	1.9
Longitud máxima de puntaje más alto	**10**	**40**	**15**	**20**	**25**
Límite inferior seleccionado	**0.5**	**0.5**	**0.2**	**0.3**	**0.7**
Límite superior seleccionado	**1.2**	**1.4**	**1.4**	**1.5**	**2.0**

2.1.2.2 Corpus lematizados con y sin información del diccionario

Una vez determinados los rangos de ratios permitidos y las longitudes máximas más adecuadas, se lematizaron los textos de forma automática mediante dos herramientas:

a. *Para shipibo-konibo:* Se utilizó el lematizador desarrollado por José Pereira, el cual fue entrenado con 460 palabras y tiene una precisión del 67% (Pereira, 2016).
b. *Para español:* Se usó la herramienta TreeTagger para español (Centrum für Informations und Sprachverarbeitung).

Ya que el lematizador para español reemplaza los números por la etiqueta *@card@,* se aplicó una regla de transformación sobre el texto ya lematizado en shipibo, con la finalidad de que ambos tengan el mismo formato. Luego, se eliminaron las duplas *[@card@,@card@]* que contienen únicamente un número tanto en shipibo como en español. Asimismo, cuando se presentaba más de una opción de lema para una palabra, se consideró únicamente la primera opción (esto sólo sucedió en el corpus en español).

Se realizaron pruebas en los cuatro dominios en base a los rangos de ratios y las longitudes máximas determinadas especificados en las tres últimas filas de la Tabla N°11. Para cada dominio, se consideraron tres tipos de corpus:

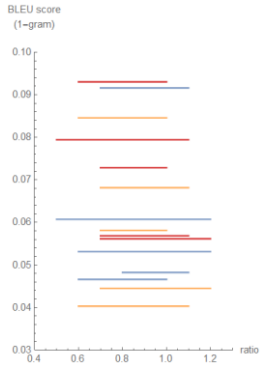
a. NL_ND: oraciones no lematizadas y sin diccionario *(estas pruebas fueron realizadas anteriormente, tal como se detalla en la sección 2.1.2.1)*
b. L_ND: oraciones lematizadas y sin diccionario
c. L_D: oraciones lematizadas y con diccionario

Los resultados con los mejores puntajes obtenidos se muestran en la Imagen N°12.

Se puede ver que en todos los dominios, salvo el religioso, se obtuvo un puntaje BLEU-1 más alto al utilizar corpus lematizados con información adicional del diccionario. Posiblemente, esto se deba a que la biblia contiene nombres propios que no pueden ser distinguidos por el lematizador, por lo que se tendrían lemas erróneos al intentar utilizar la herramienta en nombres propios. A pesar de esto, el dominio religioso nuevamente fue el dio los mejores resultados. Finalmente, los parámetros de los 5 ensayos con los puntajes más altos en cada dominio se muestran en la Tabla N°12.

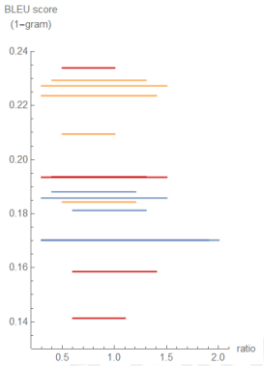
Imagen N°12. Gráficos de puntaje BLEU-1 vs. ratio de oración para los diferentes dominios. Debajo de cada gráfico se indica el puntaje máximo obtenido en cada caso. Se muestran únicamente los 5 mejores resultados de cada uno de los casos mencionados (NL_ND, L_ND y L_D).

DOMINIO DICCIONARIO



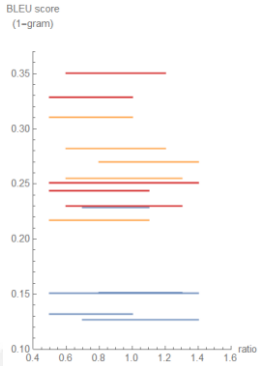
NL_ND	0.09162
L_ND	0.08459
L_D	**0.09305**

DOMINIO EDUCATIVO



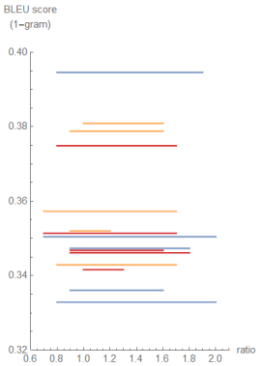
NL_ND	0.18809
L_ND	0.22936
L_D	**0.23392**

DOMINIO LEGAL



NL_ND	0.22866
L_ND	0.31051
L_D	**0.35043**

DOMINIO RELIGIOSO



NL_ND	**0.39460**
L_ND	0.38091
L_D	0.37489

LEYENDA		
	NL_ND	oraciones no lematizadas y sin diccionario
	L_ND	oraciones lematizadas y sin diccionario
	L_D	oraciones lematizadas y con diccionario

Tabla N°12. Top 5 de resultados obtenidos en cada uno de los dominios al filtrar el corpus utilizando con distintas combinaciones de parámetros

	PARÁMETROS DEL CORPUS		BLEU-1
DOMINIO EDUCATIVO *(longitud máxima 20)*	L_D	0.5-1.0	**0.23392**
	L_ND	0.4-1.3	0.22936
	L_ND	0.3-1.5	0.22729
	L_ND	0.3-1.4	0.22364
	L_ND	0.5-1.0	0.22362
DOMINIO LEGAL *(longitud máxima 40)*	L_D	0.6-1.2	**0.35043**
	L_D	0.5-1.0	0.32868
	L_ND	0.5-1.0	0.31051
	L_ND	0.6-1.2	0.28212
	L_ND	0.8-1.4	0.26995
DOMINIO DICCIONARIO *(longitud máxima 10)*	L_D	0.6-1.0	**0.09305**
	NL_ND	0.7-1.1	0.09162
	L_ND	0.6-1.0	0.08469
	L_D	0.5-1.1	0.07944
	L_D	0.7-1.0	0.07287
DOMINIO RELIGIOSO *(longitud máxima 25)*	NL_ND	0.8-1.9	**0.39460**
	L_ND	1.0-1.6	0.38091
	L_ND	0.9-1.6	0.37880
	L_D	0.8-1.7	0.37489
	L_ND	0.7-1.7	0.35726

2.2 Variación del modelo de lenguaje

En MOSES, es posible variar el modelo de lenguaje utilizado, por lo que se realizaron pruebas para distintos números de secuencia de palabras: *2-gram, 3-gram, 4-gram y 5-gram*). Las longitudes máximas permitidas permanecieron constantes para cada dominio y utilizaron 5 variaciones de corpus para cada dominio, en base a los mejores resultados obtenidos en la sección anterior (combinación de ratio de oración y transformación de oraciones con los puntajes más altos). Los resultados se muestran en la Imagen N°13 y las combinaciones de parámetros mencionadas se resumen en la Tabla N°13.

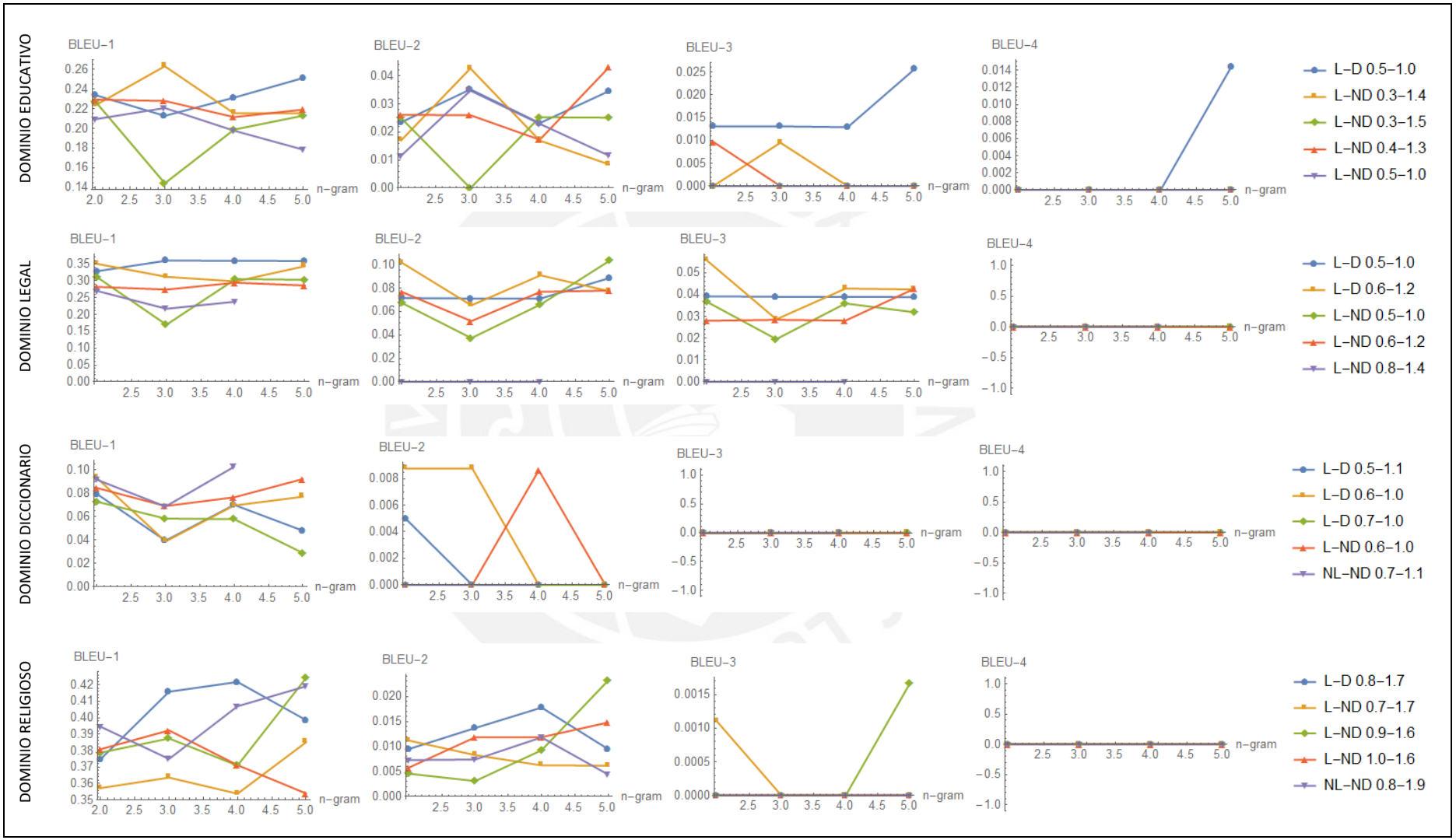
De la Imagen N°13 se observa lo siguiente:

a. *Dominio educativo:* En general, los mejores resultados se obtienen utilizando modelos de lenguaje de 3 o 5 palabras y la combinación de parámetros que da mejores resultados es L-ND_0.3-1.4 (línea amarilla).
b. *Dominio legal:* No se observa una tendencia clara al analizar únicamente la métrica BLEU-1. Sin embargo, la combinación L-D_0.6-1.2 (línea amarilla) es la que tiene puntajes más altos, predominantemente cuando trabaja con modelos de lenguaje de 2 palabras.
c. *Dominio diccionario:* Igual que en dominio legal, se determinó que los puntajes más altos se obtienen al trabajar con modelos de lenguaje de 2 palabras, particularmente para el caso L-D_0.6-1.0 (línea amarilla).
d. *Dominio religioso:* En este caso sí se observa una tendencia. Los resultados tienden a mejorar cuando se trabaja con modelos de lenguaje de más palabras. En particular, al trabajar con modelos de 5 palabras, la combinación L-ND_0.9-1.6 (línea verde) es la que tiene puntajes más altos.

Tabla N°13. Resumen de combinaciones de parámetros con los resultados más altos en cada dominio

DOMINIO	PARÁMETROS
Educativo	L-ND_0.3-1.4
Legal	L-D_0.6-1.2
Diccionario	L-D_0.6-1.0
Religioso	L-ND_0.9-1.6

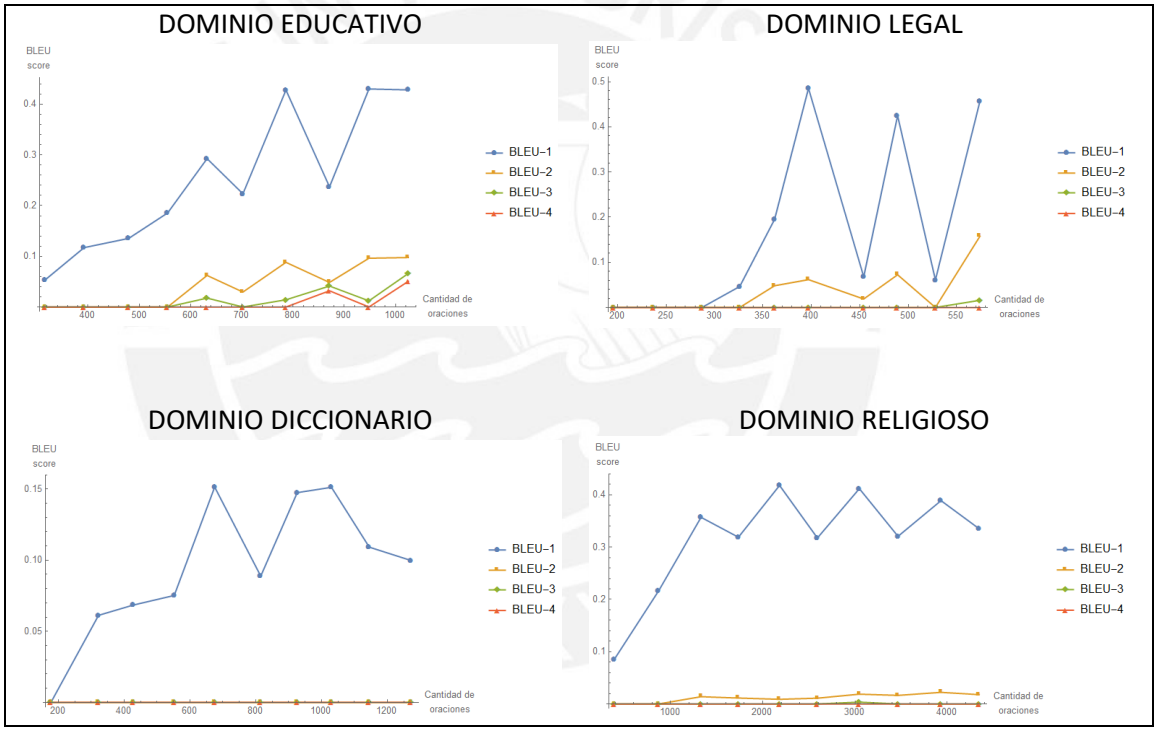
Imagen N°13. Gráficos de puntaje BLEU-n vs. número de secuencia de palabras con las que se entrenó el modelo de lenguaje (n-gram)



2.3 Variación de la longitud del corpus

Dado que la cantidad de oraciones traducidas y alineadas en shipibo-konibo y español hasta Diciembre del 2016 es bastante limitada, se realizaron pruebas para analizar si el puntaje de la métrica BLEU mejora al tener mayor número de oraciones. Con este fin, se generaron 10 sets de prueba para cada dominio, en los que la única variable era el tamaño del corpus (cantidad de oraciones), a partir de los cuales se crearon automáticamente los textos de entrenamiento, afinación y prueba. Las oraciones de los sets de prueba fueron seleccionadas aleatoriamente en base a los parámetros de la Tabla N°13. Los resultados se muestran en la Imagen N°14.

Imagen N°14. Gráficos de puntaje BLEU vs. cantidad de oraciones de entrenamiento (proporcional al tamaño del corpus). Se muestran los resultados para BLEU-1 BLEU-2, BLEU-3 BLEU-4 y BLEU-5.



En el dominio educativo, se observa claramente una relación entre el tamaño del corpus y el puntaje BLEU obtenido. A mayor número de oraciones en el corpus, el resultado de la métrica es mayor. Entonces podemos suponer que, para este dominio, se obtienen mejores resultados cuando el corpus es más grande.

Por otro lado, en el dominio legal no se puede determinar si los puntajes BLEU-1 mejoran o no al aumentar el tamaño del corpus, pero sí se evidencia un incremento en el puntaje BLEU-2. Sin embargo, el tamaño de este corpus es considerablemente reducido, por lo que se pueden obtener resultados diferentes si se contase con más oraciones.

Los resultados BLEU-1 mejoran inicialmente tanto en los dominios Diccionario como religioso, pero luego, este incremento en el puntaje se detiene. En el caso del dominio religioso, los puntajes BLEU-2 aumentan levemente conforme se incrementa el tamaño del corpus.

En general, se esperaba que los resultados mejorasen al aumentar el tamaño del corpus. Sin embargo, hay que considerar que los textos utilizados son muy pequeños en comparación con los de la literatura. En particular, el corpus más extenso en shipibo-konibo (perteneciente al dominio religioso) cuenta con 12 801 oraciones alineadas, mientras que en el corpus vasco-español descrito en la revisión sistemática tiene 50 000 oraciones (Mayor, 2011). Debido a la corta extensión de los textos utilizados, no se puede realizar un análisis detallado sobre la relación entre el tamaño del corpus y el puntaje BLEU obtenido pues, al remover oraciones aleatoriamente, es probable que otras características del corpus (además del número de oraciones) cambien.

2.4 Pruebas finales

Ya que en los ensayos anteriores se realizaron únicamente traducciones del shipibo konibo al español, en la última etapa se entrenó al traductor para que realice traducciones del español al shipibo-konibo en los 4 dominios, utilizando nuevamente los parámetros de la Tabla N°10.

Asimismo, se tradujeron todas las oraciones de los 4 corpus mediante la un traductor automático basado en reglas, con el objetivo de poder determinar de qué manera se integrarán el traductor automático basado en reglas (RBMT) y aquel basado en métodos estadísticos (SMT) en la plataforma de traducción híbrido a desarrollar en el futuro. La implementación del RBMT realizada consiste en la aplicación de reglas generadas a partir de las definiciones del diccionario en los corpus. Se tradujeron tanto oraciones lematizadas como no lematizadas, a fin de poder comparar el aporte de los lematizadores en cada uno de los dominios.

Adicionalmente, se analizaron los resultados obtenidos al validar el traductor utilizando un texto de prueba igual al texto de entrenamiento (test=train) en comparación con los resultados obtenidos anteriormente al utilizar un texto de prueba distinto al de entrenamiento (test≠train). Se decidió realizar esta prueba debido a que cuando se realizan experimentos con un traductor basado en métodos estadísticos no se pueden traducir palabras que no se encuentran en el corpus de entrenamiento, lo que disminuye el puntaje BLEU obtenido. Se realizaron estos ensayos adicionales en los dominios educativo y religioso. Los otros dos dominios no fueron considerados debido a que los textos seleccionados en etapas anteriores eran aquellos que contenían información del diccionario en los textos de entrenamiento. Entonces, al colocar las palabras del diccionario en los

textos de prueba, era probable que no se encontrase la traducción adecuada para estas, ya que cada palabra puede aparecer más de una vez en el corpus de entrenamiento, por lo que tiene múltiples traducciones posibles.

Los resultados obtenidos se presentan en la Tabla N°14. Se observa que en todos los dominios el traductor basado en reglas es el que tiene puntajes más altos. Además, el dominio que presenta mejores resultados al utilizar SMT es el religioso, probablemente debido a que cuenta con el corpus más extenso. Por otro lado, el dominio diccionario es el que tiene puntajes más altos para RBMT, lo cual era de esperarse, pues las oraciones que conforman el corpus fueron escritas para ejemplificar los términos que se utilizaron para elaborar las reglas del traductor basado en reglas.

Tabla N°14. Resumen de resultados obtenidos. El puntaje BLEU es igual al promedio de los puntajes BLEU-1 a BLEU-4.

			BLEU	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
DOMINIO EDUCATIVO	RBMT	NL ship → esp	0.19913	4.19935	0.27593	0.10088	0.07185
		NL esp → ship	0.10139	1.47746	0.12349	0.03508	0.01651
		L ship → esp	0.17842	**4.90328**	0.22084	0.08861	0.07044
		L esp → ship	0.10307	2.13011	0.08780	0.03339	0.01808
	SMT 3-gram L –ND 0.3-1.4	test≠train ship → esp	0.00000	0.26326	0.04270	0.00951	0.00000
		test≠train esp → ship	0.00000	0.09233	0.00000	0.00000	0.00000
		test=train ship → esp	0.00000	0.42922	0.07433	0.02037	0.00840
		test=train esp → ship	0.00000	0.12363	0.00811	0.00309	0.00000
DOMINIO LEGAL	RBMT	NL ship → esp	1.64015	**10.49021**	2.83941	1.19095	0.44495
		NL esp → ship	0.81418	5.42107	1.34362	0.42895	0.14064
		L ship → esp	0.40197	9.23676	0.69074	0.14121	0.05439
		L esp → ship	0.20028	3.29606	0.29049	0.07487	0.02244
	SMT 2-gram L-D 0.6-1.2	test≠train ship → esp	0.00000	0.35043	0.10156	0.05556	0.00000
		test≠train esp → ship	0.00000	0.17248	0.05837	0.03238	0.00000
DOMINIO DICCIONARIO	RBMT	NL ship → esp	0.22420	13.41279	1.19349	0.20878	0.00789
		NL esp → ship	0.00000	7.25792	0.29802	0.01275	0.00000
		L ship → esp	0.54502	**22.52372**	2.44822	0.44300	0.02912
		L esp → ship	0.00000	10.64922	0.48104	0.01939	0.00000
	SMT 2-gram L-D 0.6-1.0	test≠train ship → esp	0.00000	0.09305	0.00875	0.00000	0.00000
		test≠train esp → ship	0.00000	0.05235	0.00000	0.00000	0.00000
DOMINIO RELIGIOSO	RBMT	NL ship → esp	0.05455	4.58579	0.09501	0.00896	0.00227
		NL esp → ship	0.02659	2.56608	0.03486	0.00325	0.00175
		L ship → esp	0.06279	**6.51053**	0.15759	0.00893	0.00170
		L esp → ship	0.02746	3.72475	0.04898	0.00271	0.00117
	SMT 5-gram L-ND 0.9-1.6	test≠train ship → esp	0.00000	*0.42412*	0.02313	0.00167	0.00000
		test≠train esp → ship	0.00000	*0.44586*	0.01306	0.00000	0.00000
		test=train ship → esp	0.00000	0.42802	0.01543	0.00000	0.00000
		test=train esp → ship	0.00000	0.37761	0.01633	0.00152	0.00035

También, al comparar los traductores de shipibo-konibo a español con el de español a shipibo-konibo, se encuentra que el primero tiene mejor desempeño en los 4 dominios analizados. Asimismo, se observa que al utilizar RBMT, los puntajes BLEU son mejores cuando los textos están lematizados en los dominios educativo, religioso y Diccionario. Lo opuesto ocurre en el dominio legal, aunque la diferencia es de sólo 0.02 puntos en el caso de BLEU-1.

Finalmente, en base a los resultados obtenidos, se propone que la plataforma de traducción híbrida a desarrollar por el Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada y lingüistas de la PUCP se base en reglas, pues se necesitan corpus muy extensos para que un traductor automático basado en corpus sea viable. Entonces, dado que los recursos de tiempo y dinero son limitados, se recomienda que el traductor utilice reglas para las etapas de segmentación y decodificación, y métodos estadísticos en la etapa de reordenamiento.

V. RESULTADOS ALCANZADOS

1. Se recolectaron 4 documentos digitales que contienen oraciones en shipibo-konibo y su traducción en español:
    a. Diccionario Shipibo-Castellano
    b. Ley Forestal y de Fauna Silvestre N° 29763
    c. Ley de Lenguas Indígenas u Originarias N° 29735
    d. Hoja de ruta o resumen del plan de consulta previa del proyecto Hidrovía Amazónica

   Además, se almacenaron los textos de la Biblia Católica tanto en español como en shipibo-konibo.

2. Se generaron documentos MS Word y MS Excel con formato adecuado para que un traductor bilingüe experto pueda realizar la traducción de texto por frases (de español a shipibo-konibo y viceversa). Específicamente, se tradujeron del shipibo-konibo al español 15 documentos MS Word con aproximadamente 100 oraciones cada uno. Asimismo, se generaron 12 archivos MS Excel con texto en español, de los cuales se obtuvieron 663 oraciones. Los textos traducidos pertenecen al dominio educativo.

3. Se generaron 8 documentos en formato TXT que contienen oraciones alineadas en shipibo-konibo y español. Los textos pertenecen a 4 dominios diferentes: Diccionario, educativo, religioso y legal. Cada uno de estos dominios tiene dos archivos de texto: uno en shipino-konibo y otro en español.

4. Se elaboraron múltiples programas en *Wolfram Mathematica 11*, los cuales permiten limpiar, extraer y ordenar oraciones alineadas provenientes de un archivo de texto.

5. Se desarrolló un programa en *Wolfram Mathematica 11,* el cual tiene como función alinear automáticamente un texto por frases en base a signos de puntuación.

6. Se generaron 6 archivos de texto para cada dominio: 2 archivos con textos de entrenamiento (tanto en shipibo-konibo como en español), 2 con textos de afinación y 2 con textos de prueba. Cada uno de estos pares de archivos alineados contienen frases que fueron seleccionadas semi-automáticamente de manera que puedan entrenar traductor automático basado en corpus y dar los mejores resultados posibles dada la longitud de los corpus obtenidos.

7. Se generaron 6 archivos de texto *lematizado* para cada dominio: 2 archivos con textos de entrenamiento (tanto en shipibo-konibo como en español), 2 con textos de afinación y 2 con textos de prueba. Estos textos fueron generados automáticamente en base a los archivos descritos en el inciso anterior.

8. Se determinaron los valores más adecuados de proporción de palabras en shipibo-konibo y español en una misma oración, así como el número óptimo de palabras por oración para procesar el corpus alineado antes de utilizarlo para entrenar el traductor.

9. Para cada uno de los dominios, se determinó si es mejor trabajar con oraciones lematizadas o no lematizadas. Asimismo, se analizó si se tienen mejores resultados al incorporar las definiciones del diccionario a los corpus.

10. Se implementó un traductor automático basado en métodos estadísticos que permite la traducción de frases y palabras de shipibo-konibo a español y viceversa. Dicho traductor tiene como base la plataforma MOSES de traducción automática. Las herramientas que utiliza MOSES podrán integrarse a una plataforma de traducción híbrida.

11. Se generó un repositorio de textos traducidos del español al shipibo-konibo y viceversa. Además, para cada texto traducido, se generó un archivo en el que se indica el puntaje BLEU obtenido.

12. Se elaboraron múltiples gráficos para visualizar el puntaje BLEU obtenido al realizar pruebas de traducción variando el modelo de lenguaje y la longitud del corpus, entre otros parámetros.

13. Se generaron diversas tablas para evaluar el desempeño del traductor automático basado en corpus. Éstas, en combinación con los gráficos mencionados en el inciso anterior, permitieron identificar oportunidades para optimizar el traductor.

14. Se identificó la etapa en la que el módulo estadístico apoyará a la plataforma de traducción híbrida: éste debe servir como apoyo a una plataforma de traducción híbrida guiada por reglas.

VI. CONCLUSIONES

El objetivo general de este trabajo consistió *en desarrollar un traductor automático basado en métodos estadísticos que sirva de apoyo una plataforma de software de traducción automática de texto entre el español y el shipibo-konibo.* Para poder cumplir con este objetivo, fue necesario generar corpus paralelos en español y shipibo-konibo, pues este recurso es necesario para poder entrenar un traductor automático basado en métodos estadísticos.

Entonces, se recolectaron, tradujeron y alinearon textos en shipibo-konibo en español en 4 dominios diferentes (religioso, educativo, legal y Diccionario). El corpus del dominio religioso es el más extenso (13 257 oraciones), seguido del dominio Diccionario (5 143 oraciones), el dominio educativo (2 120) y el dominio legal (1 142). Cabe resaltar que 1 457 oraciones del dominio educativo fueron traducidas del shipibo-konibo al español por el maestro bilingüe Juan Agustín, mientras que las 663 restantes fueron traducidas del español al shipibo-konibo por estudiantes de educación bilingüe.

Por otro lado, se desarrolló un módulo de software para la traducción automática de texto entre el español y el shipibo-konibo y viceversa, para lo cual se utilizó la plataforma MOSES. En base a los experimentos realizados en las primeras etapas, se determinó que es necesario filtrar las oraciones de los corpus paralelos antes de utilizarlas para entrenar un traductor automático basado en métodos estadísticos, ya que este procedimiento mejora el desempeño del traductor, incluso si se recorta el tamaño del corpus. Asimismo, si se traducen más oraciones del dominio educativo para generar un corpus más extenso, es posible mejorar los resultados obtenidos.

Finalmente, se realizó la traducción de textos del español al shipibo-konibo y viceversa para luego validarlos utilizando la métrica BLEU. Los resultados revelan que puede obtenerse un mejor traductor automático basado en métodos estadísticos si se procesa el corpus con herramientas de apoyo como lematizadores. Además, ya que los resultados al utilizar un RBMT son mejores que los del SMT y que se tiene una cantidad limitada de oraciones para poder entrenar el traductor automático basado en métodos estadísticos**,** se determina que plataforma híbrida de traducción automática a implementar debe ser guiada por reglas.

VII. TRABAJOS FUTUROS

Para mejorar los resultados obtenidos al entrenar el SMT, podría desarrollarse un corrector ortográfico de shipibo-konibo que estandarice el texto escrito. Una misma palabra puede aparecer de manera distinta en el texto de entrenamiento y en el de prueba debido a errores tipográficos, por lo que no sería traducida. En este caso, un corrector ortográfico sería de utilidad, pues subsanaría la palabra escrita incorrectamente.

Por otro lado, es posible incorporar etiquetas de clases lingüísticas, lo que mejoraría el orden de palabras del texto de salida, de forma que se generen textos más fluidos. Asimismo, se podría mejorar la precisión el lematizador al utilizar más palabras para su entrenamiento, con lo que se generarían textos lematizados de mayor calidad.

Si se combinan todas las herramientas disponibles para realizar una primera versión de la plataforma de traducción híbrida, se podrían traducir textos de español a shipibo-konibo automáticamente (y viceversa), los cuales serían corregidos por un traductor bilingüe. Esto aceleraría el proceso de traducción, pues el traductor humano no tendría que redactar los textos desde cero. Con estos textos, se puede reentrenar al traductor automático y de esta manera mejorar los resultados obtenidos. Se sugiere hacer especial énfasis en la traducción de textos del dominio educativo, ya que el traductor automático híbrido a implementar sería utilizado principalmente para textos educativos.

Los modelos obtenidos al entrenar el SMT pueden ser utilizados para obtener mejores traducciones en la plataforma híbrida. Por ejemplo, el modelo de lenguaje mejora la fluidez del texto

generado, por lo que se sugiere que el SMT sea usado en la etapa de reordenamiento de frases. Asimismo, ya que una palabra puede tener traducciones diferentes dependiendo del contexto, el SMT puede ayudar a determinar cuál será la mejor traducción en casos específicos.

ANEXO N°1

EJEMPLO DE DOCUMENTO DE TRADUCCIÓN DEL SHIPIBO-KONIBO AL ESPAÑOL

N°	ORACIÓN A TRADUCIR	ORACIÓN TRADUCIDA
1	PEOKIN YOIYA	Primera lectura
2	Bake ainbo, bake benbo	La niña, el niño
3	Nato kirika TSANAS janeya riki 4 baritiaya bakebo shinanxon axona	Este libro TSANAS ha sido elaborado pensando en los niños de cuatro años de edad
4	TSANAS riki maxkoshoko yoina, mari jisa kikin ishto itan koshi, jawen raniriki ranshintani kikin soi, jiwi kinimeran oxai, paon piai itan mesko yobinbo jawekiaxon banai, bake pikoketia jaki yoina itan koshonara bake chixoyamatani aniai ja riki tsanas	El TSANAS es un animalito pequeño, parecido al añuje, muy veloz y fuerte, su pelo es amarillento y brillante; además, muy fino, duerme en el hueco de un árbol, come el fruto de pan del árbol, además de otros frutos y luego siembra sus semillas; también, cuando nace un bebé, el curandero puede hacer un ícaro de tsanas y así el bebé nunca tendrá diarrea y crecerá sano.
5	Nato kirika riki axeamisbobetan tsinkixon aka, mesko jawekibo jainoa oinax mia axetikopi, min anibaon itan mibe axeaibaonribi mia akinti atipanke, jaskaribiakin min axeamisninribi mia akintibo jake jainxon bena shinanbo min bibotikopi	Este libro se ha elaborado entre todos los docentes para que puedan informarse y guiarse tanto sus padres como sus compañeros de clase. Este libro puede ser utilizado por tus profesores como guía y, así, abrir discusiones para ampliar tus conocimientos.
…		
95	Baken paranta pei choshi jain napotai bochoai kaman, moa aka pikoke chopa joni	El niño mete las hojas secas de plátano hasta llenarlo, ahora ya está listo el muñeco de trapo.
96	Titan jan napota xaba kexekin xepoai	La mama termina de coser la abertura.
97	Baken pancha bero taxnanxon axonai jawen kexa itan jawen rekin, jonin bemanakeska banetiakin	Con las semillas de leucena, el niño pega su boca y su oreja para que quede listo el rostro del muñeco.
98	Titan xopo yoman boo axonai	La mamá le prepara el cabello con la lana
99	Bakebaon titan akinaton jawen chopa axonai	Con la ayuda de las mamás de los niños, se les prepara sus ropas
100	(Ainbo iamax benbon saweti)	(Vestido de mujer o de varón)

BIBLIOGRAFÍA

Aaron, L. F., & Lidia, S. (2012). LEPOR: A robust evaluation metric for machine translation with augmented factors. *24th International Conference on Computational Linguistics.*

Adelaar, W. V. (2011). *Estudios sobre lenguas andinas y amazónicas. Homenaje a Rodolfo Cerrón-Palomino.* Lima: PUCP.

Al-Onaizan, Y. C. (1999). Statistical Machine Translation. *Final Report, JHU Summer Workshop*, *30.*

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65-72.

Bismark, P. V. (2016). Aspectos morfosintácticos de la relativización en shipibo-konibo (pano). *Boletim do Museu Paraense Emílio Goeldi Ciências Humanas, 1(1)*, 123-134.

Cambridge SMT System. (2016, Mayo). Retrieved Diciembre 2016, from http://ucam-smt.github.io/tutorial/intro.html

Centrum für Informations und Sprachverarbeitung. (n.d.). *TreeTagger - a part-of-speech tagger for many languages*. Retrieved from http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Chen, L. L. (2014). A statistical method for translating chinese into under-resourced minority languages. *Source of the DocumentCommunications in Computer and Information Science*, 49-60.

Costa-Jussa, M. R. (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech & Language, 32*(1), 3-10.

Dimeo, C. (2014). *Building an Automatic Translation System from English to Scots.* University of Edinburgh.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Morgan Kaufmann Publishers Inc.*

Google. (n.d.). *Google Translate*. Retrieved from http://translate.google.com/about/intl/en_ALL/

Heafield, K. (n.d.). *Estimating Large Language Models with KenLM*. Retrieved Diciembre 2016, from http://kheafield.com/code/kenlm/estimation/

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013, Agosto). Scalable Modified Kneser-Ney Language Model Estimation. *ACL (2)*, 690-696.

HLT Machine Translation. (2013). *IRSTLM*. Retrieved Diciembre 2016, from https://hlt-mt.fbk.eu/technologies/irstlm

Hoja de ruta o resumen del plan de consulta previa del proyecto Hidrovía Amazónica. (n.d.). Perú.

Hutchins, J. (2007). *Machine translation: A concise history. Computer aided translation: Theory and practice.*

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT summit, 5*.

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge: Cambridge University Press.

Koehn, P. (2016). *MOSES: Statistical Machine Translation System. User manual and code guide.* University of Edinburgh.

Mayor, A. A. (2011). Document Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, 53-82.

Mermer, C. K. (2017). The TÜbİTAK-UEKAE statistical machine translation system for IWSLT 2007. *IWSLT*, 176-179.

Ministerio de Agricultura. (n.d.). Jakoni jiwibo itan yoinabo kai aki teetinin ley N° 29763. Perú.

Ministerio de Agricultura y Riego. (n.d.). Ley forestal y de fauna silvestre N° 29763. Perú.

Ministerio de Cultura. (2014). Ley N° 29735 Jaskaxon non joibo itan non shinanbo, keyomarestima kopi jato onamabokin non Perúkamea joibo iki ika Ley. Perú.

Ministerio de Cultura. (n.d.). *Base de Datos de Pueblos Indígenas y Originarios: Shipibo-konibo*. Retrieved Mayo 1, 2016, from http://bdpi.cultura.gob.pe/pueblo/shipibo-konibo

Ministerio de Educación. (2012). *Dirección General de Educación Intercultural, Bilingüe y Rural*. Retrieved from http://www.minedu.gob.pe/digeibir/

Ministerio de Educación. (2014). *Axeti kirika Tsanas - 4 Baritiayabaona.* Lima, Perú: Consorcio Corporación Gráfica Navarrete S.A.

Ministerio de Educación. (2015, Junio 16). *Minedu oficializa alfabetos de 24 lenguas originarias a ser utilizados por todas las entidades públicas*. Retrieved Abril 2016, from Ministerio de Educación: http://www.minedu.gob.pe/n/noticia.php?id=33082

Ministerio de Educación. (2016). *Sesiones de Aprendizaje 2016*. Retrieved 2016, from http://www.minedu.gob.pe/rutas-del-aprendizaje/sesiones2016/primaria.php

NTT Communication Science Labs. (2014). *Rank-based Intuitive Bilingual Evaluation Score*. Retrieved from http://www.kecl.ntt.co.jp/icl/lirg/ribes/

Och, F. J. (n.d.). *GIZA++: Training of statistical translation models*. Retrieved 2016, from http://www.fjoch.com/giza-training-of-statistical-translation-models.html

Pa, W. T. (2016). A Study of Statistical Machine Translation Methods for under Resourced Languages. *Procedia Computer Science*, 250-257.

Papineni, K. R. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, (pp. 311-318). Philadelphia.

Pereira, J. (2016). *Implementación de un lematizador para una lengua de escasos recursos: caso shipibo-konibo.* Lima.

Pérez, A. A.-I. (2012). EuskoParl: A speech and text Spanish-Basque parallel corpus. *13th Annual Conference of the International Speech Communication Association 2012*, 2359-2362.

Pinnis, M., & Skadiņš, R. (2012). MT adaptation for under-resourced domains-what works and what not. *Frontiers in Artificial Intelligence and Applications*, 176-184.

Pinnis, M., Skadiņa, I., & Vasiļjevs, A. (2013). Domain adaptation in statistical machine translation using comparable corpora: Case study for English Latvian IT localisation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 224-235.

PUCP. (n.d.). *Red Internacional de Estudios Interculturales (RIDEI)*. Retrieved from Centro de Estudios Filosóficos: http://cef.pucp.edu.pe/investigacion/red-internacional-de-estudios-interculturales-ridei/

Rank-based Intuitive Bilingual Evaluation Score. (n.d.). Retrieved from http://www.kecl.ntt.co.jp/icl/lirg/ribes/

Salami, S., Shamsfard, M., & Khadivi, S. (2016). Phrase-boundary model for statistical machine translation. *Computer Speech and Language*, 13-27.

San Pablo España. (n.d.). *Biblia Latinoamericana*. Retrieved 2016, from http://www.sanpablo.es/biblia-latinoamericana

Skadiņa, I. L.-P. (2012). Linguistically motivated evaluation of english-latvian statistical machine translation. *Frontiers in Artificial Intelligence and Applications*, 221-229.

SRI International. (2016, Noviembre). *Downloading and Building SRILM*. Retrieved Dieicmbre 2016, from http://www.speech.sri.com/projects/srilm/download.html

Teenoxon jaskakin noa yokakanti shinanbo nato Idrovia Amazonica ikainko. (n.d.).

Vogel, S. (2003). Using noisy bilingual data for statistical machine translation. In A. f. Linguistics (Ed.), *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2.*

Wise, M. R. (Ed.). (1993). *Diccionario Shipibo-Castellano.* Yarinacocha, Pucallpa, Perú: Instituto lingüístico de verano.

World Bibles. (n.d.). *The Bible in Shipibo-Conibo*. Retrieved 2016, from http://worldbibles.org/language_detail/eng/shp/Shipibo-Conibo

Wycliffe Bible Translators, Inc. (2012). *Diossen joi jatíxonbi onanti joi (SHPNTPO)*. Retrieved 2016, from https://www.bible.com/es/versions/673-shpntpo-diossen-joi-jatixonbi-onanti-joi

Zariquiey, R. (2006). Reinterpretación fonológica de los préstamos léxicos de base hispana en la lengua shipibo-conibo. *Boletín de la Academia Peruana de la Lengua, 41*.

Zariquiey, R. (2011). Aproximación dialectológica a la lengua cashibo-cacataibo (pano). *Revista Lexis, 35 (1)*.