

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:

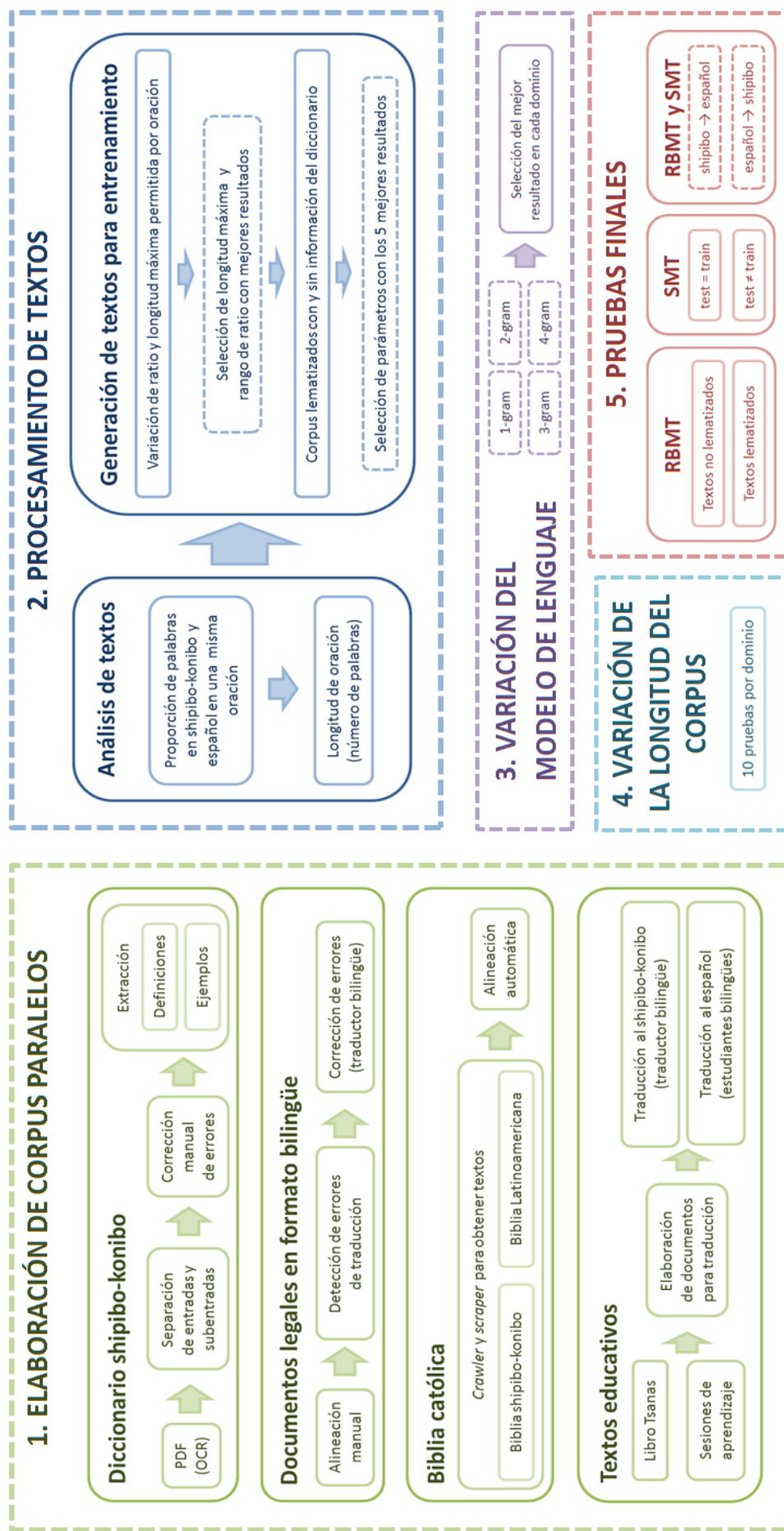
1. Elaboración de corpus paralelos
2. Procesamiento de textos
3. Variación del número de secuencia de palabras (o variación del *modelo de lenguaje*)
4. Variación de la longitud del corpus
5. Pruebas finales

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	1	cancán s. cáncaman : piña <Báqueshoco
2	cancán cócomaara óshiai. Cuando damos	2	cancán cócomaara óshiai. Cuando damos
3	de comer piña a un bebé, adelgaza.>	3	de comer piña a un bebé, adelgaza.>
4	baríncancan [del ship.	4	baríncancan (del ship.
5	<i>bári, barín sol + cancán piña]</i>	5	<i>bári, batín sol + cancán pifia]</i>
6	: *ayupa	6	: °ayupa
7	pásacancan [del ship.	7	pásacancan (del ship.
8	<i>pása, elem. no ident. + can-</i>	8	<i>pása, elem. no ident. + cancán</i>
9	<i>cán piña] : especie de</i>	9	<i>pifia] : especie de</i>
10	piña con espinas en el	10	piña con espinas en el
11	borde de sus hojas	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: <i>ábion</i>), prefijo (ej: <i>ra-</i>) o sufijo (ej: <i>-que</i>)
2	Variante del término introductorio	Categoría de la variante + variante
3	Clase	Clase de la palabra. Ejemplos: sustantivo (<i>s.</i>), pronombre (<i>pron.</i>), adjetivo (<i>adj.</i>),...
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: [<i>del ship.</i>]
7	Definición	Definición o sentido equivalente en castellano. Puede ser una sola definición (: <i>definición</i>) o más de una (<i>1: definición_ 2: definición_2 ...</i>)
8	Nota de uso	Nota o comentario acerca de la definición. Presenta la siguiente estructura: <i>-Úsase ...</i>
9	Ilustración verbal de la definición	Palabra, frase u oración que se presenta como ejemplo. Tiene la siguiente estructura: <i><shipibo . español></i>
10	Subentrada	Tiene la misma estructura que la parte principal
11	Referencias	Tienen la siguiente estructura: <i>Véase...</i>
12	Párrafo de sinónimos	Sinónimos en mayúsculas e información adicional. Tiene la siguiente estructura: <i>sinón. ...</i>

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:

- Ley Forestal y de Fauna Silvestre N° 29763* (Ministerio de Agricultura) (Ministerio de Agricultura y Riego)
- 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)
- 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 Baritiyabaona* (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	
<i>Axeti kirika Tsanas: 4 Baritiayabaona</i>	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:

$$\text{ratio}_{\text{oración}} = \frac{n^{\circ}\text{ship}}{n^{\circ}\text{esp}} = \frac{\text{número de palabras en shipibo - konibo}}{\text{número de palabras en español}}$$

Se espera que en la mayoría de casos las oraciones tengan más palabras en español que en shipibo-konibo ($\text{ratio} < 1$) 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.

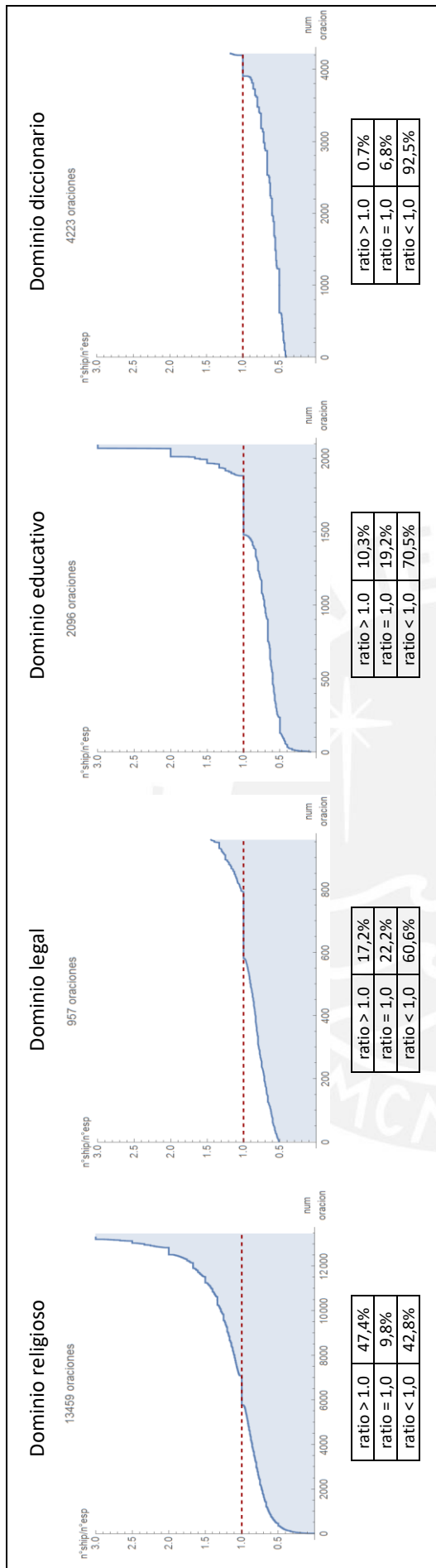
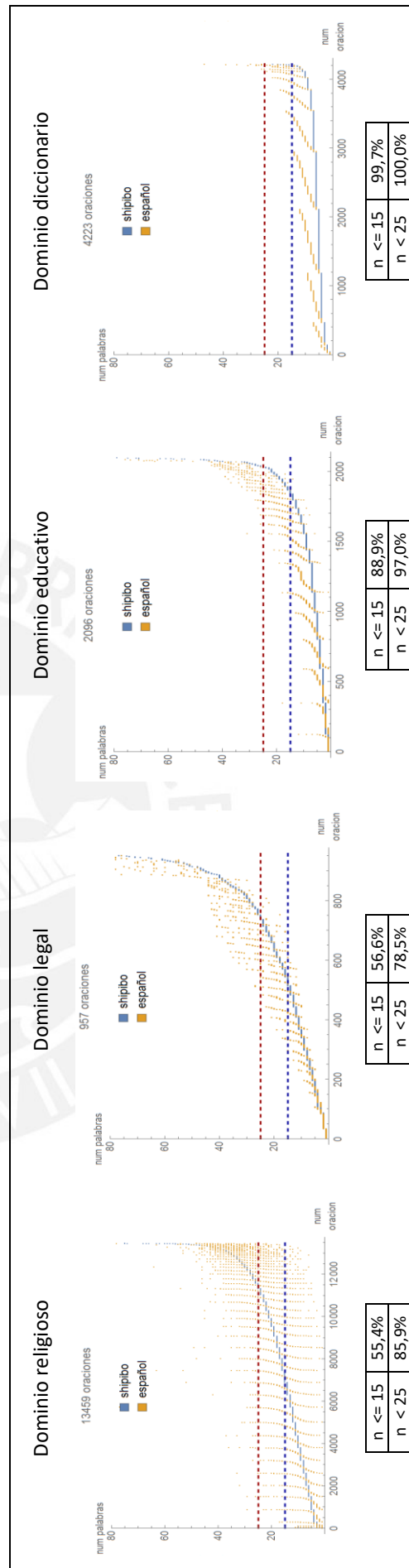


Imagen N°8. Número de palabras en shipibo-konibo y en español (por oración). La línea horizontal azul corresponde a un número de palabras en shipibo-konibo igual a 15 y la roja a 25.



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
<p><i>rínsira jáke tibítoain nóriti amirikain jaínoax orópain ja riki míxo jisá jaboen jína máxkoxoko jawen ráni riki kextó jawen bíchi riki kopóra</i></p> <p>(22 palabras)</p>	<p><i>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</i></p> <p>(30 palabras)</p>

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:

- Oraciones diferentes tienen un número diferente de palabras (**longitud** de oración)
- La proporción de palabras en shipibo-konibo y español varía (**ratio** de oración)
- Las palabras de las oraciones pueden ser **lematizadas** automáticamente
- 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 estos 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 shipibokonibo 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 shipibokonibo. 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 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.

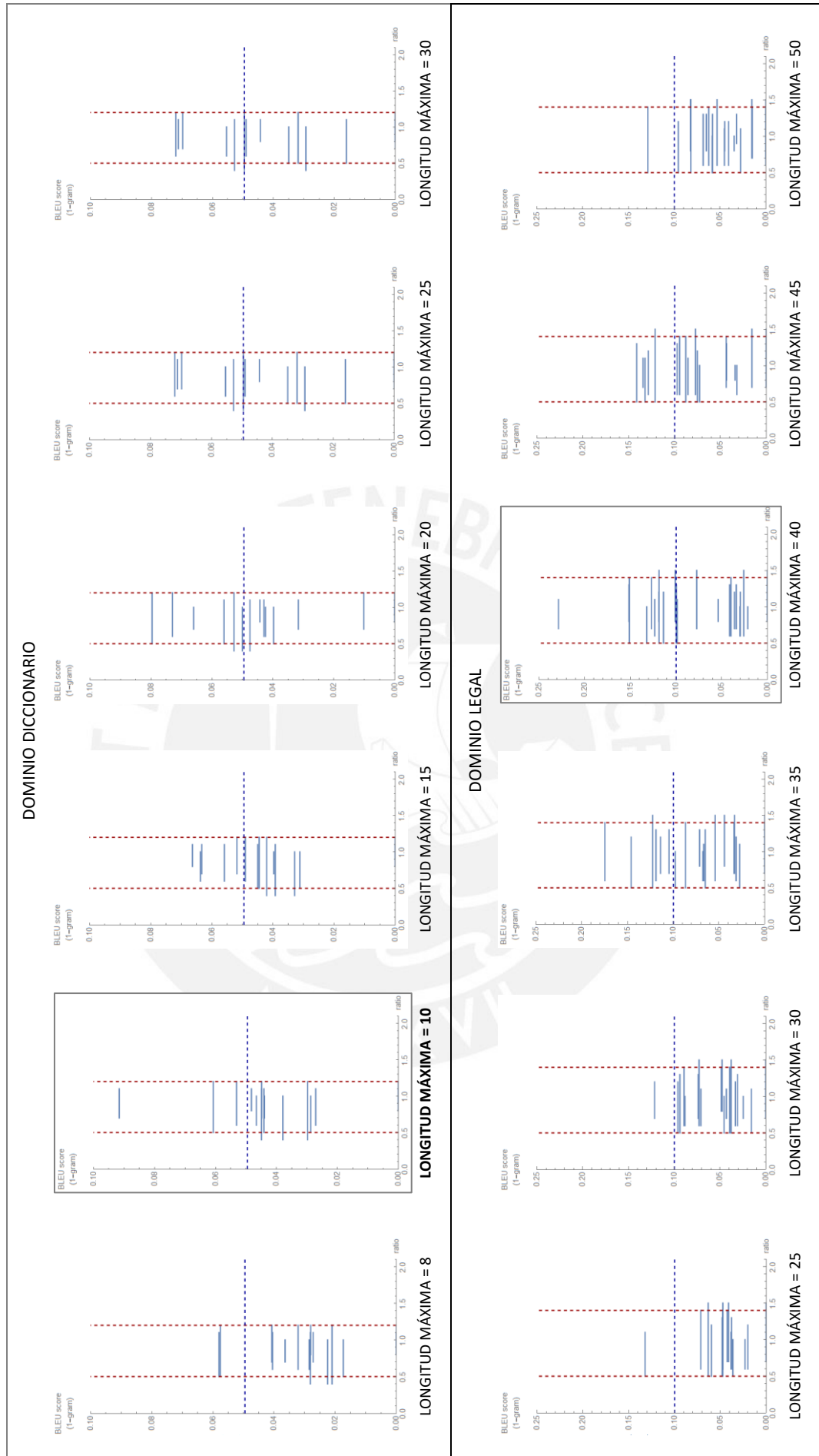


Imagen N°10. Gráficos de Puntaje BLEU (1-gram) vs. ratio de oración para diferentes límites de longitud en los dominios educativos. La línea horizontal azul punteada corresponde al resultado obtenido sin aplicar ningún filtro y las líneas rojas al rango de ratio seleccionado para realizar los siguientes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

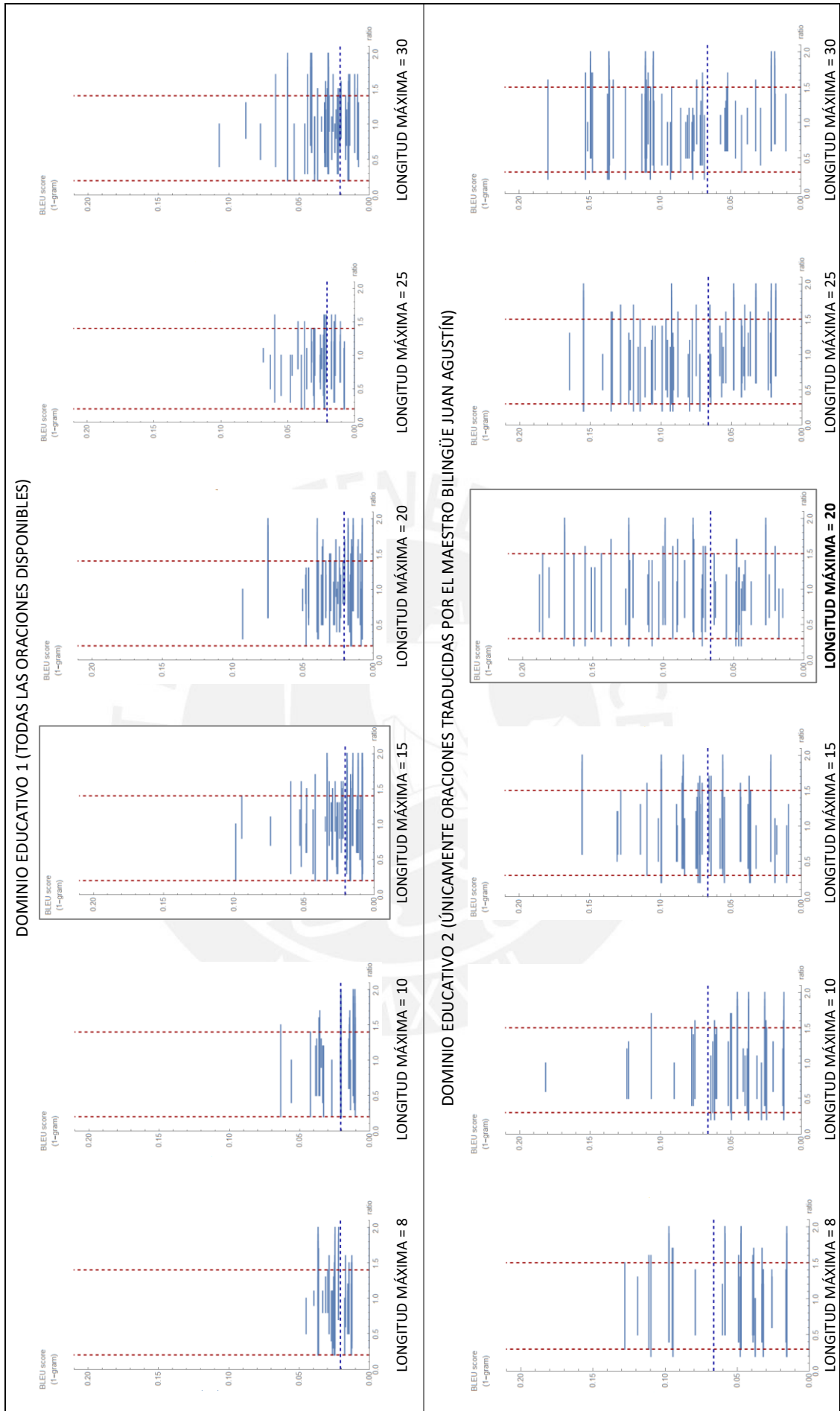


Imagen N°11. Gráficos de Puntaje BLEU (1-gram) vs. ratio de oración para diferentes límites de longitud en el dominio religioso. La línea horizontal azul punteada corresponde al resultado obtenido sin aplicar ningún filtro por longitud y las líneas rojas al rango de ratio seleccionado para realizar los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

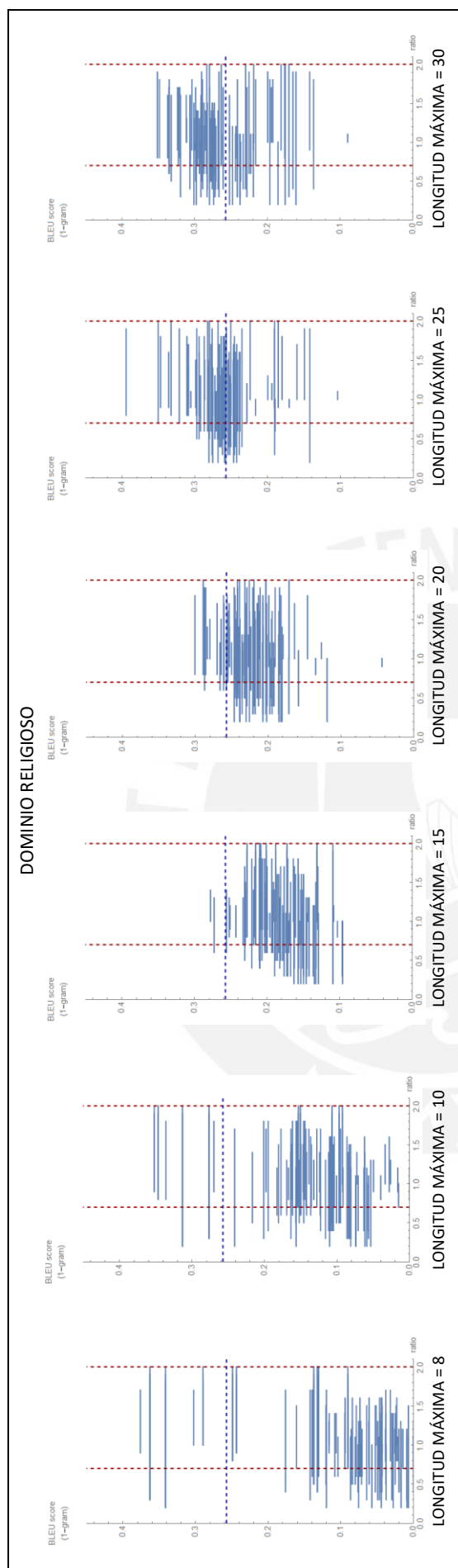


Tabla N°11. Tabla resumen de los resultados obtenidos al variar la longitud máxima permitida y el ratio en cinco dominios diferentes.

	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).

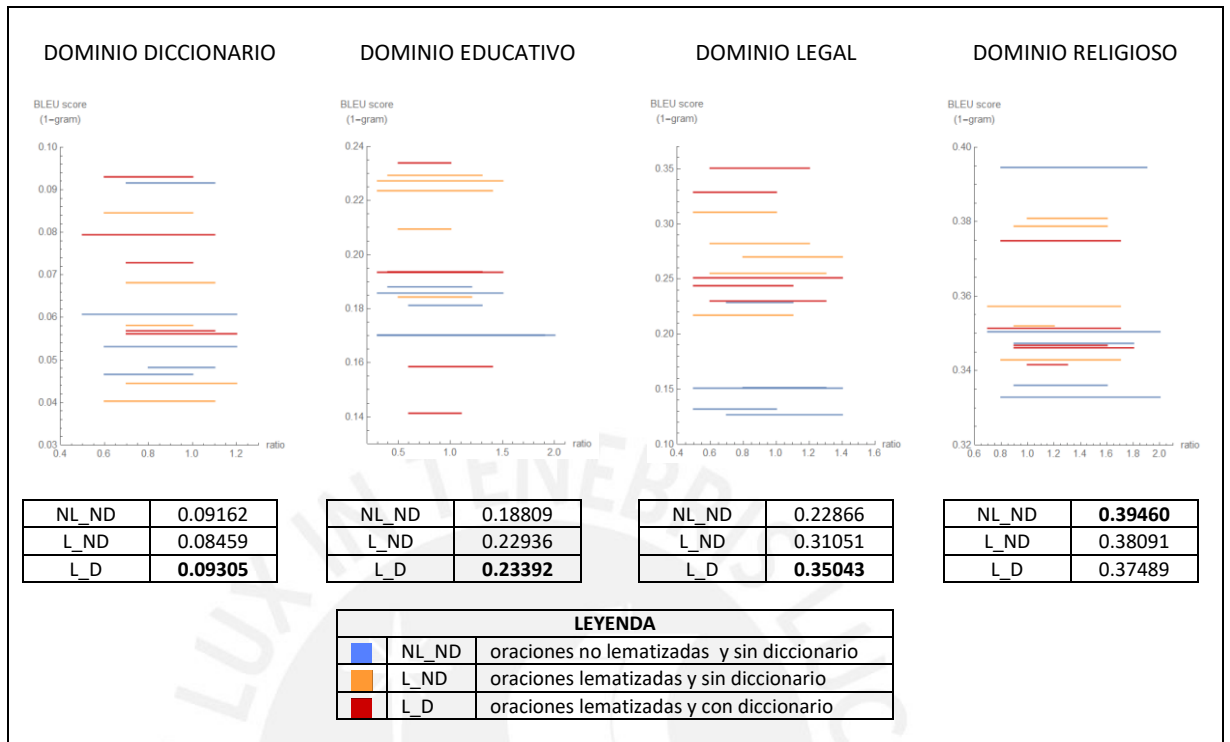


Tabla N°12. Top 5 de resultados obtenidos en cada uno de los dominios al filtrar el corpus utilizando con distintas combinaciones de parámetros

	PARÁMETROS DEL CORPUS		BLEU-1
DOMINIO EDUCATIVO <i>(longitud máxima 20)</i>	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 <i>(longitud máxima 40)</i>	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 <i>(longitud máxima 10)</i>	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 <i>(longitud máxima 25)</i>	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.

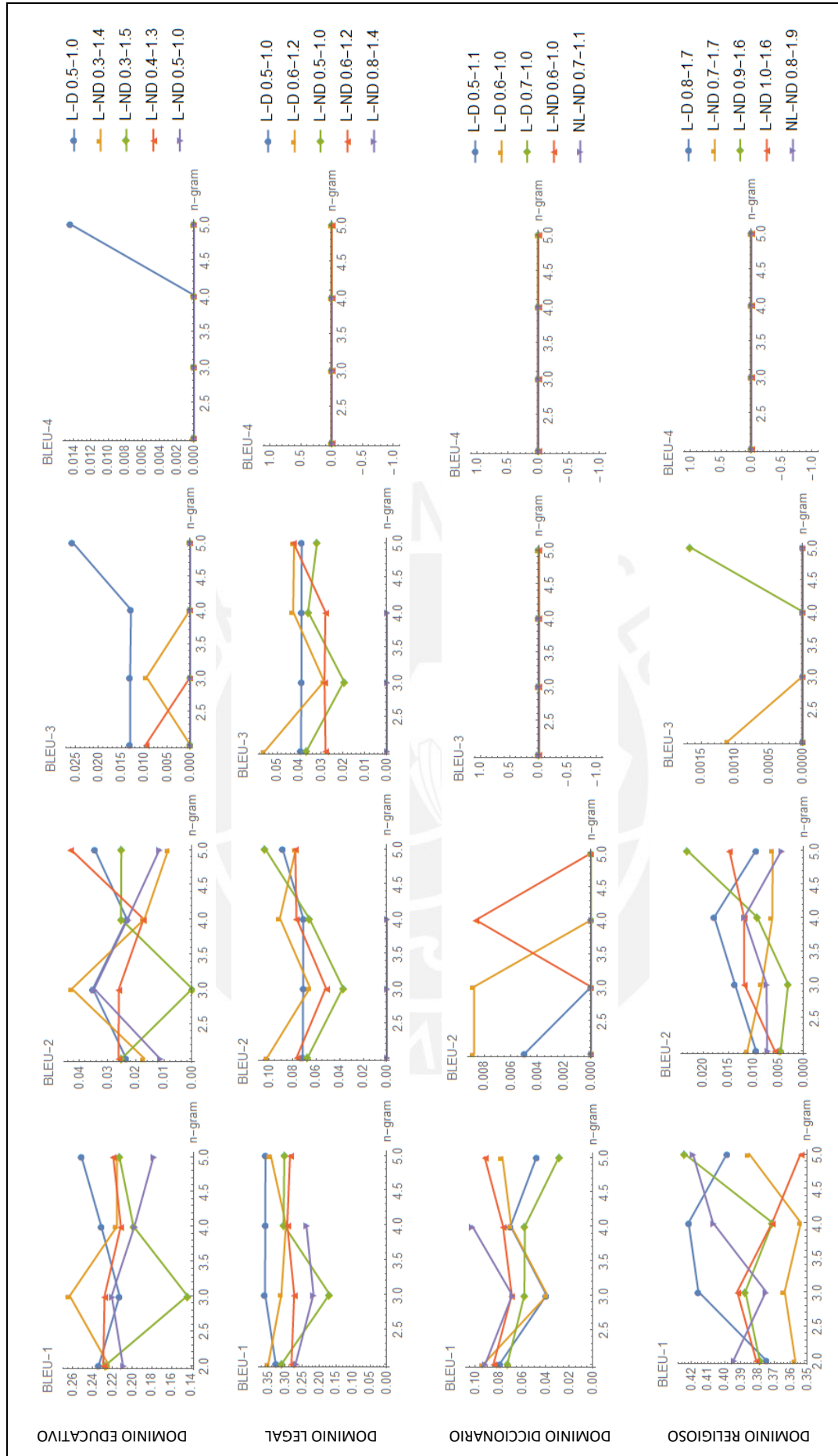
De la Imagen N°13 se observa lo siguiente:

- 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

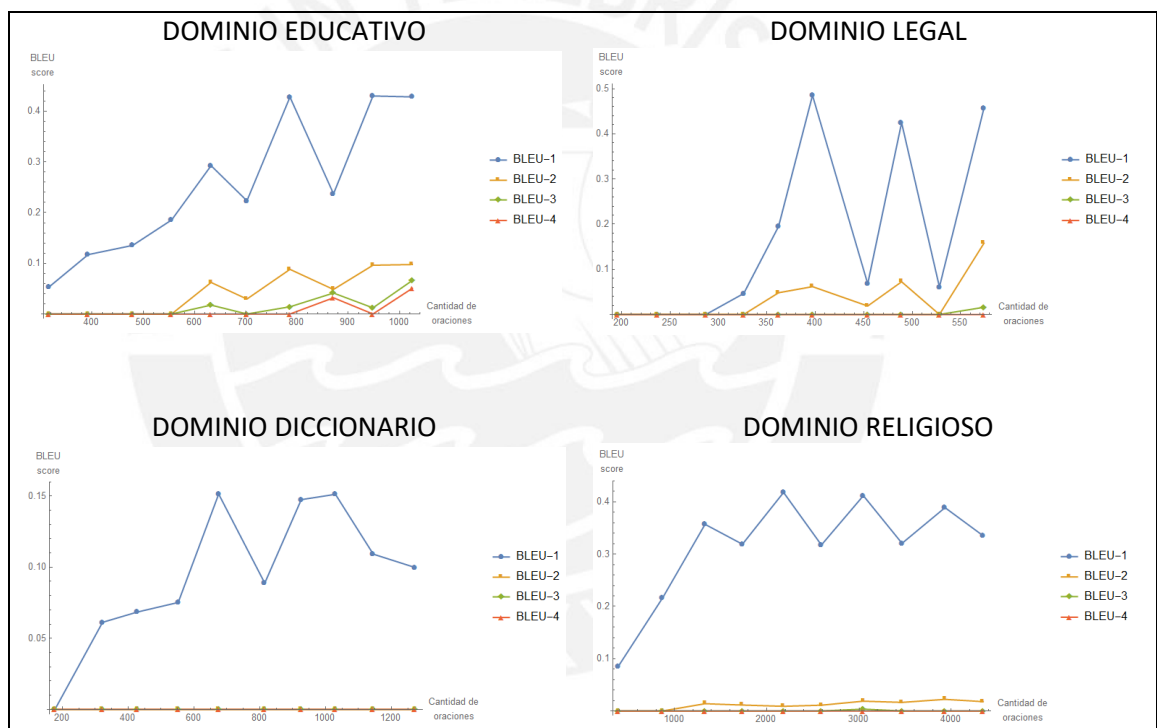
Imagen N°13. Gráficos de puntaje BLEU-n vs. número de secuencia de palabras con las que se entrenó el modelo de lenguaje (n-gram)



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 encontrara 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	<i>0.42412</i>	0.02313	0.00167	0.00000
		test≠train esp → ship	0.00000	<i>0.44586</i>	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 AmazónicaAdemás, 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 shipibo-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 a 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 del 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 axeibaonribi 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 chopajoni	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 mamá 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 akinton jawen chopaxonai	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 Document Communications 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 TÛbĪTAK-UEKAE statistical machine translation system for IWSLT 2007. *IWSLT*, 176-179.
- Ministerio de Agricultura. (n.d.). Jakoni jiwibo itan yoinabo kai aki teetin 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 Baritiyabaona*. 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., & Vasiljevs, 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 Diciembre 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 jatixonbi 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).