UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENIERÍAS
PROGRAMA DE INGENIERÍA FÍSICA

TRABAJO DE GRADO:

IMPLEMENTACIÓN DE UN SISTEMA DE TRADUCCIÓN AUTOMÁTICA BASADO EN MODELOS ESTADÍSTICOS PARA LA TRADUCCIÓN DE LA LENGUA DE SEÑAS COLOMBIANA AL ESPAÑOL

DIANA CAROLINA LÓPEZ BUSTAMANTE
ESTEFANÍA MARÍN ARCILA

Director:
PhD JULIÁN DAVID ECHEVERRY CORREA

DICIEMBRE 2017

Proyecto de grado presentado como requisito final para optar por el título de Ingeniería Física

IMPLEMENTACIÓN DE UN SISTEMA DE TRADUCCIÓN AUTOMÁTICA BASADO EN MODELOS ESTADÍSTICOS PARA LA TRADUCCIÓN DE LA LENGUA DE SEÑAS COLOMBIANA AL ESPAÑOL

Este trabajo de grado hace parte del proyecto "Metodología para el Reconocimiento y la Traducción de Señas Aisladas en la Lengua de Señas Colombiana Utilizando Técnicas de Visión por Computador" avalado por la Vicerrectoría de Investigaciones, Innovación y Extensión de la Universidad Tecnológica de Pereira

DIANA CAROLINA LÓPEZ BUSTAMANTE
ESTEFANÍA MARÍN ARCILA

Director:
PhD JULIÁN DAVID ECHEVERRY CORREA

DICIEMBRE 2017

Agradecemos a nuestras familias por su apoyo, su motivación y por creer siempre en nosotras. Pero en especial, por su amor incondicional.

Agradecemos al grupo "Los de Séptimo" por su amistad brindada.

Y desde nuestras más sinceras intenciones, agradecemos a Julian David Echeverry Correa porque más que un gran director es una excelente persona.

Diana Carolina López Bustamante - Estefanía Marín Arcila

Índice general

Índice de figuras

Índice de cuadros

Lista de acrónimos

- ALPAC:*Automatic Language Processing Advisory Committee* - Comité Asesor para el Procesamiento Automático del Lenguaje
- BLEU: *BiLingual Evaluation Understudy*
- DANE: Departamento Administrativo Nacional de Estadísticas
- FENASCOL: Federación Nacional de Sordos de Colombia
- IBM:International Business Machines
- ICAL: Fundación Para el Niño Sordo
- INSOR: Instituto Nacional para Sordos
- LM: *Language Model* - Modelo de Lenguaje
- LS: Lenguas de Signos
- LSC: Lengua de Señas Colombiana
- MT: *Machine Translation* - Traducción Automática
- OMS: Organización Mundial de la Salud
- RM: *Reordering Model* - Modelo de Reordenamiento
- SER: *Sentence Error Rate* - Porcentaje de Frases Erróneas
- SMT: *Statistical Machine Translation* - Traducción Automática Estadística
- TM: *Translation Model* - Modelo de Traducción
- UTP: Universidad Tecnológica de Pereira
- WER: *Word Error Rate* - Porcentaje de Palabras Erróneas
- WFD: Federación Mundial de Sordos

RESUMEN

Los humanos, en condición de seres sociales, necesitan comunicarse. La comunicación es el intercambio de ideas mediante un código de conocimiento mutuo. No siempre la comunicación se da de forma exitosa, existen condiciones que limitan el proceso comunicativo tales como: la codificación del mensaje (idiomas o lenguas) o las habilidades limitadas de transmisión o emisión, como es el caso de las discapacidades sensoriales. Las discapacidades sensoriales son la ceguera, sordera y dificultad del habla [1]. La disminución gradual o total de la capacidad auditiva es uno de los factores que influye en el aprendizaje de la lengua, lo que conlleva a buscar formas alternativas de comunicación como lo es la lengua de señas.

Para el caso de los sordos colombianos, la Lengua de Señas Colombiana (LSC) es una lengua transmitida por medio del movimiento de las extremidades superiores y representada de forma escrita por glosas. La LSC es un código de comunicación para la población no oyente, pero actualmente, existen dificultades para entablar una comunicación efectiva con las personas oyentes, debido a que un alto porcentaje de la población colombiana no conoce o no sabe la LSC. Dicha problemática trasciende en temas como lo es el acceso a la educación, ya que en los centros educativos no se cuenta con suficientes intérpretes que abarquen todos los puntos y aulas de la instalación.

El presente trabajo abordó una parte de la problemática, por medio de un prototipo de un sistema de MT. El prototipo facilita la comunicación entre personas sordas y oyentes en un solo sentido; en un punto particular de las instalaciones de un centro de educación superior, específicamente, la ventanilla de información del Centro de Registro y Control de la Universidad Tecnológica de Pereira (UTP) El sistema de traducción elaborado en este trabajo, aplica técnicas de procesamiento de lenguaje natural, específicamente traducción automática estadística (SMT) basada en frases. Además de hacer uso de herramientas y *toolkits* tales como Moses, Giza++ , Kenlm, entre otras.

Para el desarrollo del sistema de traducción, como primera medida se construyó un *corpus* paralelo, actualmente formado por 517 frases en glosas con su respectiva traducción en español. Todas las frases del *corpus* son perteneciente a un solo dominio: el punto de información de Registro y Control de la UTP

El *corpus* paralelo fue dividido aleatoriamente en 3 partes, creando 3 diferentes *subcorpus* paralelos: entrenamiento, validación y evaluación. A su vez, el procedimiento anterior se

realizó 10 veces. Es decir, quedando 10 carpetas formada cada una por un *corpus* paralelo de entrenamiento, validación y evaluación.

Posteriormente se realizó el procesamiento de texto a los *corpus* de cada una de las 10 carpetas, se generó el modelo de lenguaje y el modelo de traducción. Se hizo la evaluación del sistema por medio de la aplicación de 2 métricas de evaluación. Obtenidos los resultados, se procedió a realizar el ajuste correspondiente al sistema con el fin de mejorar la traducción.

En el proceso de validación se realizó 4 experimentos diferentes. Al ser evaluados los experimentos, los resultados obtenidos por la métrica BLEU están en promedio de 26,46 % ±0,53 para sistemas de traducción con modelos de lenguaje de trigramas y del 25,40 % ±0,53 para sistemas de traducción con modelos de lenguaje de bigramas. Para el caso de la métrica WER están en promedio de 9,55 % ±0,36 para sistemas de traducción con modelos de lenguaje de trigramas y del 9,67 % ±0,36 para sistemas de traducción con modelos de lenguaje de bigramas. El resultado del proyecto fue muy satisfactorio.

Capítulo 1

INTRODUCCIÓN

Las personas con discapacidad son aquellas que tienen alguna restricción en la capacidad de realizar una actividad de manera normal para el ser humano; esto como consecuencia de una malformación congénita o un desarrollo en el transcurso de la vida. Las discapacidades se dividen en físicas, sensoriales, psicológicas, intelectuales y psiquiátricas. La discapacidad sensorial comprende la ceguera, sordera y dificultad del habla [1].

La sordera es la pérdida de la audición en uno o ambos oídos, bien sea total o parcial. Es importante diferenciar entre personas sordas prelocutivas, perilocutivas y postlocutivas, es decir, las personas que quedaron sordas antes, durante o después de adquirir la lengua oral o escrita. Una persona prelocutiva es aquella que pierde la capacidad para oir antes de adquirir la lengua, es decir, antes de los 2 años de edad; las personas prelocutivas tienden a padecer de mudez [2]. Una persona perilocutiva es aquella que adquirió la discapacidad durante el aprendizaje de la lengua, aproximadamente entre los 2 y 4 años. Por último, una persona poslocutiva es aquella que pierde la capacidad de oír después de adquirir la lengua, es decir en una edad posterior a los 3 años.

Toda la población es vulnerable a la perdida de audición debido a un proceso natural llamado presbiacusia o perdida de la audición gradual, que inicia a partir de los 20 años de edad y se hace más evidente a partir de los 50 años de edad [3]. La presbiacusia puede ser acelerada por el uso de medicamentos agresivos para el nervio auditivo o por una larga exposición a espacios ruidosos.

Según la Organización Mundial de la Salud (OMS), hay 360 millones de personas que padecen pérdida de audición discapacitante [4] lo que equivale al 5 % aproximadamente de la población mundial. Según el censo poblacional del Departamento Administrativo Nacional de Estadísticas (DANE) del 2005, en Colombia el 1,1 % de la población tiene limitaciones para oír; en Risaralda se estima que ese porcentaje corresponde al 1,2 % , mientras que en Pereira, es del 1,3 % [5]. La Universidad Tecnológica de Pereira (UTP) tiene aproximadamente 21 personas con este tipo de discapacidad entre estudiantes de pregrado, posgrado y egresados

[1].

A nivel mundial, existen organizaciones tales como la Federación Mundial de Sordos (WFD) [2] que es una organización no gubernamental y que cuenta con 133 miembros de diferentes países. La WFD trabaja por las oportunidades de igualdad y promueve el uso correcto de la lengua de señas como garantía de acceso a la educación.

En Colombia, entidades como el Instituto Nacional para Sordos (INSOR) [3], la Fundación Para el Niño Sordo (ICAL) [4], la Fundación DIME [5] y la Federación Nacional de Sordos de Colombia (FENASCOL) [6] trabajan por el mejoramiento de la calidad de vida de las personas sordas mediante la protección de sus derechos, la creación de políticas acorde a la necesidad de dicha población y el desarrollo e implementación de acciones dentro del sector educativo.

Cabe resaltar, que las mayores dificultades presentadas entre las personas sordas son a la hora de relacionarse con otras personas en un entorno de oyentes. Esto es debido a la diferencia del código lingüístico, lo que complica las relaciones familiares y sociales, acceso a medios de transporte y salud, desempeño en el mundo laboral y en especial, la formación durante el periodo educativo. Una de las causas por las cuales las personas sordas no acceden o abandonan el sistema educativo es porque visualizan la discapacidad como una incapacidad [6]. Esta situación de abandono escolar puede mejorarse con la utilización de audífonos, implantes cocleares, el empleo de subtítulos, el aprendizaje de la lengua de señas y otras medidas de apoyo educativo y social.

A pesar de que la UTP promueve procesos de inclusión en la educación superior con cursos de formación en la Lengua de Señas Colombiana (LSC) para toda la población universitaria, la adquisición de estas herramientas es insuficiente, lo cual dificulta la comunicación bilateral entre las personas sordas y oyentes en el dominio académico.

La traducción automática (*Machine Translation* - MT) es un área de la lingüística computacional que investiga el uso de software para traducir texto o habla de un lenguaje natural a otro. Esta se presenta como una alternativa a la comunicación entre personas sordas y oyentes. La MT se puede realizar de dos formas: basada en reglas y basada en *corpus* lingüísticos. La última, a su vez, se divide en MT por memoria y por estadística. La traducción automática estadística (*Statistical Machine Translation* - SMT) consiste en modelos estadísticos cuyos análisis se obtienen de un *corpus* de textos bilingües construido por glosas y texto en cualquier idioma.

Este trabajo hace parte de un proyecto financiado por la Vicerrectoría de Investigacio-

[1]Cifra oficial entregada por la Universidad Tecnológica de Pereira en la oficina de Vicerrectoría de Responsabilidad Social y Bienestar Universitario

[2]Página oficial de la Federación Mundial de Sordos url: https://wfdeaf.org

[3]Página oficial del Instituto Nacional para Sordos url: http://www.insor.gov.co

[4]Página oficial de la Fundación Para el Niño Sordo url: http://www.icalcolombia.org

[5]Página oficial de la Fundación DIME url: http://www.dimecolombia.org

[6]Página oficial de la Federación Nacional de Sordos de Colombia url: https://www.fenascol.org.co

nes, Innovación y Extensión de la UTP, que tiene como objetivo formular y desarrollar una metodología para el reconocimiento de señas aisladas en LSC y su traducción al castellano utilizando técnicas de visión por computador. Este trabajo aportará al proyecto en la sección de técnicas de MT más usadas dentro del área de procesamiento de lenguaje natural; iniciando con la recopilación de una base datos y finalizando con la construcción del modelo de traducción.

El fin último del proyecto de investigación es facilitar la comunicación entre estudiantes universitarios sordos y/o personas externas con los administrativos oyentes del área de Registro y Control de la UTP con el fin de que las personas sordas encuentren distintas fuentes de información.

Dicho esto, esta memoria está organizada en siete capítulos repartidos de la siguiente manera; en los capítulos 1, 2 y 3 se presenta la introducción a este proyecto de grado, el planteamiento del problema y los objetivos que se cumplieron durante el desarrollo del trabajo. En el capítulo 4 se dan los fundamentos teóricos que respalda el desarrollo y la ejecución de este trabajo. En el capítulo 5 se explica detalladamente cada uno de los pasos realizados para el cumplimiento de los objetivos y los resultados obtenidos en cada uno de ellos. En los capítulos 6 y 7 a partir de los resultados obtenidos se extraen las conclusiones y los posibles trabajos a futuro. Finalmente, se incluyen los anexos 7 y las referencias donde puede consultarse toda la bibliografía utilizada 7.

Capítulo 2

PLANTEAMIENTO DEL PROBLEMA

Un proceso de traducción completo de Lengua de Señas Colombiana (LSC) a español consta de 2 etapas. En la figura 2.1 se muestra las etapas con sus respectivas entradas y salidas.

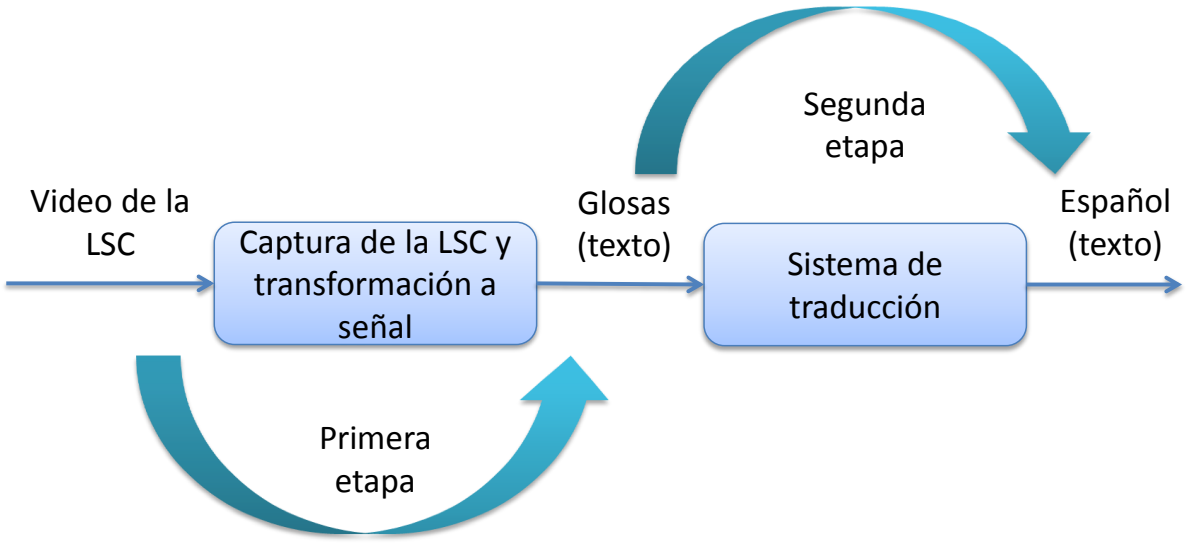


Figura 2.1: Diagrama de bloques del proceso de traducción de Lengua de Señas Colombiana a español

La primera etapa tiene como entrada un vídeo donde está representada la LSC y como salida un texto en glosa que representa la señal. Esta etapa consiste en la adquisición y el etiquetado de la LSC. La segunda etapa tiene como entrada el texto en glosa y como salida

un texto en español. Esta etapa consiste en un sistema de traducción. Para efectos de esta memoria se centró en la segunda etapa de la figura 2.1.

El primer traductor automático apareció en los años 60, traducía de ruso a inglés. Los primeros traductores, se basaban en las reglas lingüísticas características de cada idioma. Actualmente, a nivel mundial se han hecho grandes estudios acerca de como mejorar el proceso de traducción y en el desarrollo de posibles herramientas y técnicas para el mejoramiento de la traducción, por ejemplo: el proyecto TEAM de la Universidad de Pensylvania, que usa un sistema de traducción de inglés a Lenguaje de Señas Americano basado en reglas gramaticales [7]; en Sudáfrica se reutilizó el proyecto TEAM con una traducción de inglés a Lenguaje de Señas Sudafricano [8]. Seguidamente, se desarrollaron otros trabajos con enfoques estadísticos tales como el de la Universidad Politécnica de Madrid, donde se diseñó un sistema de traducción automática (*Machine Translation* - MT) denominado CONSIGNOS para comunicación bilateral, el sistema de traducción está basado en *corpus*: traducción basada en frases, en memoria y estados finitos. En este trabajo se emplearon las herramientas y *toolkits* GIZA++, Phrase-Extract, Phrase-Score, SRILM, Moses, REFX y algoritmos como GIATI. Para la evaluación se utilizaron varias métricas como: WER, BLEU y NIST [9]. En la Universidad de Túnez, se desarrolló un sistema de traducción automática estadística (*Statistical Machine Translation* - SMT), donde se utilizó modelos basados en palabras, algoritmos de IBM del 1 al 3 y el algoritmo de optimización. Se emplearon las herramientas GIZA++ y un algoritmo de concordancia de cadenas usando la distancia Jaro-Winkler y el decodificador Moses [10]. En la Universidad de Bohemia Occidental, realizaron una traducción de checo a Lenguaje Señado Checo y viceversa, en este trabajo, se usan dos decodificadores diferentes: Moses y SimPad. Se encontró que los resultados del SiMPad son comparables con los resultados obtenidos con Moses. Las métricas de evaluación fueron: SER, WER, PER y BLUE [11]. Finalmente, en la Universidad Nacional de Cheng Kung,Taiwan, se desarrolló un sistema de traducción de la Lengua de Señas China a taiwuanés utilizando el método probabilístico PCFG y el algoritmo Viterbi para elegir la mejor traducción [12].

Otros trabajos, se han centrado en emplear diferentes técnicas y algoritmos para mejorar los sistemas de MT, como: en la Escuela de Tecnología e Informática en China, se resumió cuatro métodos para la adaptación de dominio basado en: datos, modelo mixto, *corpus* monolingüe y el modelo de tema [13]. En la Universidad Politécnica de Cataluña, se propusó una nueva estrategia estadística para afrontar una de las principales fuentes de error en los sistemas de MT estocástica debido al cambio de orden de las palabra, denominado reordenamiento automático estocástico [14].

A nivel nacional, varias universidades como la Universidad Nacional [15], la Universidad Pedagógica Nacional de Colombia [16], la Universidad CES de Medellín y la Escuela de Ingeniería de Antioquia [17]; han realizado trabajos de traducción de LSC al español, donde sus estudios se centran específicamente en la adquisición de la LSC (primera etapa de la figura 2.1) teniendo como principal objetivo el reconocimiento automático y como resultado entregan el etiquetado de la LSC. Estos trabajos tuvieron en cuenta las condiciones del entorno de adquisición, donde su enfoque principal es el procesamiento de imagen, vídeo y la captura por parte de equipos especializados. Por tal motivo, esta tesis propuso un trabajo de

traducción (segunda etapa de la figura 2.1 de LSC a español donde se hace énfasis en cada uno de los pasos seguidos para llevar a cabo el proceso de traducción.

Capítulo 3

OBJETIVOS

3.1. Objetivo General

IMPLEMENTAR un sistema de traducción automática basado en modelos estadísticos para la traducción de la LSC al español en un dominio específico, puntualmente un punto de información académica de una institución de educación superior.

3.2. Objetivos Específicos

1. Estudiar distintas técnicas en el campo de los modelos estadísticos de traducción automática y analizar, a la vez, distintos sistemas que hagan uso de estas técnicas.

2. Construir una base de datos paralela (LSC- español) limitada por un dominio académico, empleando frases que puedan ser utilizadas en un punto de información de una institución de educación superior.

3. Preparar en el *corpus* paralelo, los subconjuntos de entrenamiento, validación y evaluación. Procesar los subconjuntos utilizando técnicas de procesamiento de texto, tanto para la lengua origen como la lengua destino.

4. Generar el modelo de lenguaje y el modelo de traducción para el *corpus* procesado mediante *toolkits* especializados para la construcción y aplicación de modelos estadísticos.

5. Decodificar el modelo de lenguaje y el modelo de traducción empleando un sistema de traducción estadístico.

6. Evaluar la implementación realizada haciendo uso de métodos heurísticos y métricas conocidas para la validación de sistemas de traducción automática.

Capítulo 4

ESTADO DE LA CUESTIÓN

En este capítulo se da una visión general de los fundamentos teóricos para el desarrollo de este trabajo de grado. En la sección 4.1 se menciona las principales características de las lenguas de señas, específicamente de la Lengua de Señas Colombiana (LSC). Por otro lado, en la sección 4.2 se habla acerca de los sistemas de traducción automática. La sección 4.3 describe las métricas de evaluación para la calidad del sistema de traducción y finalmente, en la sección 4.4 se realiza una breve descripción acerca del *toolkit* Moses.

4.1. Lengua de Señas Colombiana

En esta sección se abordan algunos aspectos relacionados a la LSC. Primeramente en la subsección 4.1.1 se da a conocer las generalidades de un lenguaje natural, específicamente de la LSC. En la subsección 4.1.2 se muestra algunas generalidades históricas de la LSC y finalmente, en la subsección 4.1.3 las características lingüísticas propias de esta lengua.

4.1.1. Generalidades

Las lenguas de señas, también conocidas como lenguas de signos (LS), son lenguas naturales que poseen todas las propiedades y complejidades de cualquier lengua natural oral o escrita. Se entiende por lengua natural, aquella lengua creada espontáneamente con el objetivo de generar comunicación entre humanos, a diferencia de las lenguas construidas, o lenguas artificiales, que son lenguas planeadas o diseñadas por seres humanos como por ejemplo los lenguajes de programación o lenguajes de máquina.

Las LS tienen la particularidad que son producidas por gestos y percepciones visuales. El canal de comunicación de las LS es gesto-viso-espacial, o incluso táctil, en el caso de las

personas con sordo ceguera; mientras que el lenguaje oral tiene un canal vocal-auditivo [18].

Las LS no son versiones simplificadas de las lenguas orales como se ha preconcebido y mucho menos son mímica, ya que los signos no tienen que estar directamente relacionados con una palabra [9]. Al igual que las lenguas habladas, transforman unidades sin significado (fonemas) en unidades con información semántica. Los fonemas o componentes del signo son, en este caso, la forma de la mano, la orientación de la palma, el lugar de la articulación, el movimiento y la expresión facial [19].

4.1.2. Características Históricas

El uso de señas como modo de comunicación es tan antiguo en la humanidad como las lenguas orales e incluso más; de hecho algunas tribus indígenas de Norteamérica utilizaban las señas como método de comunicación entre distintas etnias que no compartían la misma lengua. Si bien las lenguas de señas son utilizadas en gran parte por personas sordas, estas han sido y continúan siendo usadas por personas oyentes, las cuales han sido pilares fundamentales para establecer una comunicación entre personas sordas y oyentes, obrando muchas veces, como intérpretes o traductores [18].

William C. Stokoe fue un pionero en el estudio de las lenguas de señas. Entre sus legados está la clasificación de las lenguas de señas, diferenciando dos tipos: el primero es artificial y está basado en como escribir y leer una lengua oral específica, en donde los signos se refieren a letras (deletreo dactílico). El segundo es natural, como el desarrollado por los sordos en todo el mundo, el cual no está basado en ningún tipo de lengua hablada. Stokoe, en 1960 publicó la monografía "*Sign Language Structure*" (Estructura de la Lengua de Señas), en la que propone que las señas pueden ser analizadas como compuestos simultáneos de tres elementos sin significado (morfemas gestuales): una forma de la mano (queirema), una actividad de la mano (quinema) y un lugar ocupado por la mano (toponema). Este estudio lo hizo con la Lengua de Signos Americana (ASL) utilizada por sus alumnos sordos. A partir de ese momento, comenzaron a reconocerse socialmente las lenguas de signos [18].

Así como para el lenguaje oral hay diferentes idiomas en el mundo, de igual manera sucede con la lengua de señas; lo cual no la hace universal ya que para cada país hay diferentes lenguas de señas ubicadas regionalmente. Dentro de las diferentes lenguas de señas que se encuentran alrededor del mundo encontramos la Lengua de Señas Colombiana (LSC).

Según Paulina Ramírez, subdirectora de investigación del Instituto Nacional para Sordos (INSOR), los orígenes de la LSC se remontan a 1920, en un internado católico bogotano denominado Instituto de Nuestra Señora de la Sabiduría [20]. En este sitio se llevaban a cabo programas y actividades educativas para jóvenes sordos, los cuales se basaban en métodos utilizados en Francia. Estos procesos de formación fueron interrumpidos en las décadas de los sesenta y los setenta [21]. En 1957 aparece la primera asociación de sordos en Bogotá y un año después otra en Cali.

En la década de los ochenta, la comunidad sorda colombiana empezó a preocuparse en gran medida por el estudio, difusión y enseñanza de la lengua de señas. Con el transcurso de los años, se fueron conformando grupos para ejercer investigaciones referentes al lenguaje manual, dando como resultado la publicación de las primeras cartillas creadas por la Federación Nacional de Sordos de Colombia (FENASCOL) en 1997. El Instituto Nacional para Sordos (INSOR), en 1998, presentó un libro cuyo contenido recoge ensayos acerca de la historia y estructura de la LSC [22] [37]. En agosto del 2000 se inició la construcción de un diccionario de la LSC, el cual tenia inicialmente 400 señas. En el 2001 se publicó "Apuntes para una gramática" de la LSC por parte del profesor venezolano Alejandro Oviedo, en donde se ilustra las peculiaridades lingüísticas de la LSC. Finalmente, en 2006 se publicó el primer Diccionario Básico de la Lengua de Señas Colombiana con un total de 1.200 definiciones. El diccionario se publicó por parte del INSOR con apoyo de FENASCOL, el Instituto Caro y Cuervo y la Universidad del Valle. [22] [23] [37]

Es importante dar a conocer que gran parte de los signos utilizados en la Lengua de Señas Colombiana está emparentada filogenéticamente con la Lengua de Señas Francesa, y que algunos signos son muy similares a los de la Lengua de Señas Española, Americana y Salvadoreña.

A nivel normativo, en Colombia existe la promulgación de la Ley General de Educación, Ley 115 de 1994, y su decreto reglamentario, el 2082 de 1996, que establecen la integración de las poblaciones especiales al sistema educativo regular y por tanto la transformación de las Instituciones de Educación Especial [37].

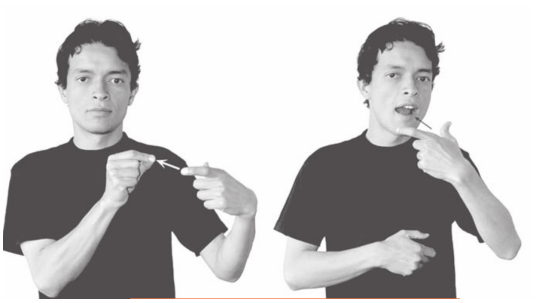
La Lengua de Señas Colombiana fue reconocida en el año 1996, con la Ley 324 donde el artículo 2 establece: "El estado colombiano reconoce la lengua de señas como propia de la comunidad sorda del país" [22].

4.1.3. Características Lingüísticas

Las características lingüísticas son las descripciones técnicas para la utilización elocuente de una lengua. A continuación, se mencionan algunas características lingüísticas de la LSC:

1. La representación escrita de las señas se expresa mediante **glosas**, estas consisten en asignar a cada seña una o varias palabras de la lengua española que representen de modo aproximado el significado base de la seña. La glosa se escribe en mayúscula sostenida. Por ejemplo, en la tabla 4.1 se hace una descripción gráfica de como se realiza la seña "anciano" y esta tiene por glosa la palabra "ANCIANO". Varias señas se pueden representar con una sola glosa y varias glosas se pueden representar con una sola seña como se muestra en la tabla 4.1 para el ejemplo de la seña de "pizza" y "hacer el amor"[1].

[1]Imágenes y ejemplos extraídos del Diccionario Básico de la Lengua de Señas Colombiana de las página

Descripción gráfica de la LSC	Representación en glosa de la LSC
	ANCIANO
	PIZZA
	HACER-EL-AMOR

Cuadro 4.1: Ejemplos de la descripción gráfica de la seña y su representación en glosa de la LSC

2. En la LSC el orden de la oración suele ser Objeto-Sujeto-Verbo, los adjetivos siempre van detrás del sujeto y los artículos no existen. Por ejemplo:

Español	Glosas de la LSC
María vende frutas	MARÍA FRUTAS VENDER
El vestido es bonito	BONITO VESTIDO
¿Cuál es la casa?	CASA ¿CUÁL?

Cuadro 4.2: Ejemplos de la sintaxis en glosa de la LSC

3. En las glosas, los lugares o personas indicadas y señaladas por el locutor se representan con palabras especiales [23]. Las cuales son:

6, 35 y 73

Palabras especiales	Significado en la glosa
INDEX	Él, ella, ellos, ustedes ahí, aquí, eso
PRO1	Yo, me
PRO1POS	Mío, mi
PRO2	Tú, te, a tí
PROPOS	Mio, tuyo, suyo, de ustedes
PRODUAL	Nosotros dos, ustedes dos, ellos dos

Cuadro 4.3: Significado de las palabras especiales utilizadas en la escritura de la glosa para representar lugares o personas señaladas por el locutor

Algunos ejemplos del uso de las palabras de la tabla 4.3 se muestran a continuación [2]:

Español	Glosas de la LSC
Yo dormía en el internado	PRO1 IR DORMIR INTERNADO
Ella tiene cinco meses de embarazo	INDEX EMBARAZO CINCO MES

Cuadro 4.4: Ejemplos de traducción del español a glosa cuando el locutor señala lugares o personas

4. En la lengua de señas no se utiliza los verbos ser/estar. Por ejemplo:

Español	Glosas de la LSC
Estaba tan feliz que soñó con la fiesta	INDEX FELIZ SOÑAR FIESTA
Yo soy médico	PRO1 MÉDICO

Cuadro 4.5: Ejemplos del uso del verbo ser y estar en español y su correspondiente traducción en glosa

5. El significado de la seña depende del uso particular del espacio, la modificación sistemática en el movimiento con el cual se produce la seña, la producción de movimientos no manuales del cuerpo y de los ojos, la expresión facial, la orientación y la posición de todo cuerpo. Por ejemplo:

[2]Ejemplos extraídos del Diccionario Básico de la Lengua de Señas Colombiana de la página 35

Glosa de la LSC	Acción	Significado
NIÑO VARÓN	Mirar hacia la derecha (producción de movimientos no manuales de los ojos)	Un niño ubicado en el lado derecho
HOMBRE FRUTA MOVER	La persona mueve la mano de abajo hacia arriba y de arriba hacia bajo mientras realiza la seña de "mover"	Un hombre baja una fruta

Cuadro 4.6: Ejemplo del significado adicional dado a la LSC por la producción de movimientos con el cuerpo

6. Para ciertas lenguas, existen morfemas cuya función es dar claridad a los objetos referidos en las frases. Los morfemas dan claridad de un objeto mediante la mención de características como el material del que están hecho, su forma o tamaño, su ubicación o disposición espacial, si se mueven o permaneces fijos, si está en estado solido o líquido, si son humanos o no, entre otros. A estos morfemas se les denomina **clasificadores**. Por ejemplo:

Significado	Frase con clasificador
Comprar café líquido	CL(dan)-COMPRAR CAFÉ
Comprar café en polvo	CL(kan)-COMPRAR CAFÉ

Cuadro 4.7: Ejemplo del uso de clasificadores en algunas lenguas

Particularmente en la LSC, los clasificadores se escriben junto al verbo y se constituyen en dos tipos: **configuraciones manuales clasificadores** que se refieren a la clase del objeto envuelto en el predicado y el segundo, **raíces de movimiento** que constituyen la raíz verbal a la cual se incorporaba la configuración manual representando el movimiento o la locación de los nombres asociados al verbo [38]. En la tabla 4.8 se muestra un ejemplo:

Glosa sin clasificador	Glosa con clasificador	Significado
HOMBRE FRUTA MOVER	CL:V(persona)DE-abajo-MOVER-A-arriba	Un hombre sube por frutas a un árbol
FRUTA MOVER	CL:4∘(Objeto esférico)DE-arriba-MOVER-A-centro	Recolectar frutas

Cuadro 4.8: Ejemplo del uso de los clasificadores en la LSC

7. En la LSC los verbos no se conjugan. Por ejemplo:

Español	Glosas de la LSC
Nosotros jugamos en el parque	PRODUAL JUGAR PARQUE
Salí tarde de la reunión	REUNIÓN TARDE SALIR

Cuadro 4.9: Ejemplos de la escritura de los verbos en glosa de la LSC

Para dar significados adicionales al verbo es necesario agregar signos como son los clasificadores (ver item anterior).

8. Para dar una explicación gramatical de la LSC es fundamental tener en cuenta tres componentes: la matriz articuladora, la matriz segmental y la matriz de rasgos no manuales [38].

   La matriz articuladora esta relacionada específicamente con las postura de la mano, sus partes móviles, su orientación y ubicación. En la figura 4.1 se muestra los puntos de referencia de las manos para la articulación de la seña [3]. En la figura 4.2 se muestra la referencia para la orientación y movimiento de las manos [4].

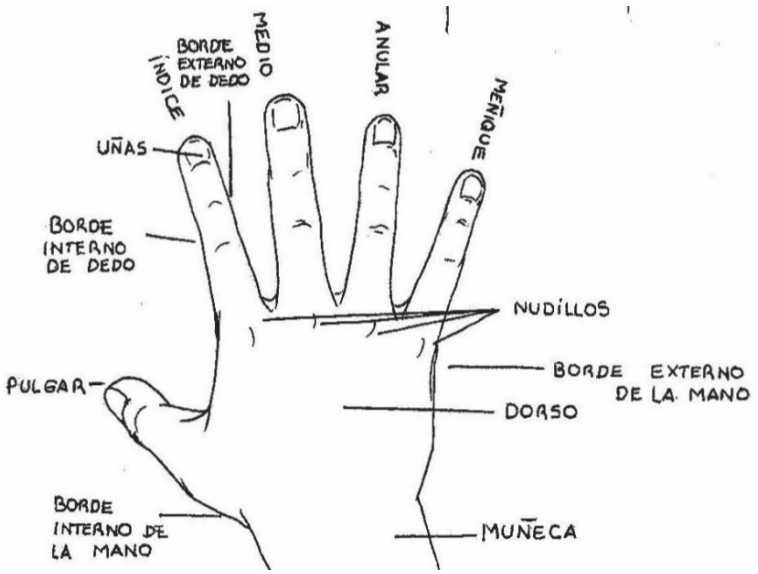


Figura 4.1: Imagen de los puntos de referencia de las manos para la articulación de la seña

[3]Imagen extraída del Diccionario Básico de la Lengua de Señas Colombiana de la página 571
[4]Imagen extraída del Diccionario Básico de la Lengua de Señas Colombiana de la página 572

Figura 4.2: Imagen de la referencia espacial para la orientación y movimiento de las manos

La matriz segmental es la encargada de identificar si dentro de una seña se producen movimientos, detenciones o transición en la detención de la o las manos. Es sumamente importante esta estructura ya que es muy parecida a la silábica en la que cada seña tiene una parte en su ejecución lo cual resulta muy semejante a las silabas de una palabra en castellano. Por ejemplo: la palabra "gracias" tiene dos silabas gra-cias, de la misma forma que en LSC posee varios momentos, la mano en forma de palma situada en el mentón seguidamente un movimiento hacia el frente el cual termina con una detención contra la palma de la otra mano como se muestra en la figura 4.3.

Figura 4.3: Representación gráfica de la palabra "gracias" en LSC

Por ultimo, encontramos la matriz de rasgos no manuales que comprende toda la información de carácter gestual los cuales no son realizados con las manos sino con la nariz,

boca, ojos, mejillas, cejas, cabeza y cuerpo; las cuales actúan como refuerzo para complementar las señas de las anteriores matrices. En la tabla 4.10 se muestra los rasgos básicos que cambian o dan sentido a la expresión neutra.

Parte del cuerpo	Postura de la parte del cuerpo
Cabeza	Adelantada, atrás, inclinada, ladeada a la izquierda, ladeada a la derecha
Ceño-cejas	Cejas arriba, ceño fruncido
Ojos	Inusualmente abiertos, semi-cerrados, cerrados
Mirada	Arriba, abajo, a un lado
Nariz	Fruncida
Lengua	Protruida, vibrando
Labios	Retraídos, abocinados, distendidos, soplando, protruidos, abiertos
Mejillas	Infladas, retraídas
Barbilla	Desplazada lateralmente, adelantada, atrás
Cuerpo	Hombros encogidos, inclinado, ladeado

Cuadro 4.10: Rasgos corporales básicos que agregan sentido a la LSC

4.2. Sistemas de Traducción Automática

En esta sección se abordan algunos aspectos relacionados con la traducción automática. Primeramente en la subsección 4.2.1 se da a conocer la historia, evolución y aplicaciones. En la subsección 4.4 se definen las clasificaciones de la traducción automática y en la subsección 4.2.3 se profundiza en un sistema de traducción, denominado traducción automática estadística. Seguidamente en la subsección 4.2.4 se describe todo lo relacionado al *corpus* continuando en la subsección con las principales restricciones de un *corpus*; por último, en la subsección 4.2.6 se menciona los principales inconvenientes de los sistemas de traducción automática.

4.2.1. Aspectos Históricos

La traducción automática, conocida como MT por sus siglas del inglés *Machine Learning*, es una de las aplicaciones más comunes del campo de la lingüística computacional.

Los inicios de la MT se remontan al año 1933, cuando el francés George Artsrouni y el ruso Petr Trojanski patentaron sus trabajos. El primero fue acerca del diseño de un dispositivo de traducción y el segundo, fue una propuesta de un método para un diccionario bilingüe automático [24]. Posteriormente, llegaron los días en que las computadoras fueron usadas en Gran Bretaña para romper el código Enigma alemán en la Segunda Guerra Mundial y en la década de los 60 se conoció el primer traductor automático que traducía de ruso a inglés.

Después de los primeros avances sobre la MT, se empezó a estudiar diferentes enfoques para traducir, desde métodos de traducción directa, métodos de transferencia, hasta métodos de interlengua. En 1956, la Universidad de Georgetown y la compañía *International Business Machines* (IBM) realizaron el experimento Georgetown-IBM, que consistió en la traducción de más de sesenta frases del ruso al inglés. El experimento se consideró un éxito, lo que dio oportunidad a que los gobiernos se preocuparan por invertir en la lingüística computacional. En Estados Unidos la mayoría de las investigaciones se centraban en la traducción del ruso al inglés por motivos políticos y militares, para el caso de Europa y Canadá, la traducción se lograba de inglés a francés por motivos culturales. En 1966, el Comité Asesor para el Procesamiento Automático del Lenguaje (*Automatic Language Processing Advisory Committee*-ALPAC), publicó un informe donde afirmó que la MT era muy costosa y que los intérpretes humanos abundaban. Tras la publicación de dicho informe, la financiación para el desarrollo e investigación de nuevas tecnologías de sistemas de traducción se dio por terminada. A pesar del declive financiero, en 1976 surgió el sistema de MT *Météo* el cual traducía pronósticos del tiempo, en 1968 se fundó *Systran* para traducir de ruso a inglés. En la década de los ochenta, se comercializaron los sistemas *Logos* (traductor de alemán a inglés y de inglés a francés), *Metal* (traductor de alemán a inglés) y los sistemas desarrollados en la Organización Panamericana de la Salud (traductor de español a inglés e inglés a español) [25].

En la mayoría de los casos, los sistemas de traducción de la época, eran bajo los modelos clásicos de traducción (traducción directa, traducción por transferencia y traducción interlengua). En 1991, la compañía International Business Machines (IBM) publicó los resultados de los experimentos de un sistema denominado *Candide*. Este sistema se basaba en métodos exclusivamente estadísticos [26]. El sistema *Candide* lo entrenaron con el *corpus Hansard* de las Actas del Parlamento Canadiense compuesta por aproximadamente tres millones de oraciones en inglés y francés.

Actualmente, la MT es una de las aplicaciones que ha generado más facilidades para la relación con otras culturas y la difusión del conocimiento. Muchos ejemplos claros son la traducción de páginas web en internet, los traductores de un idioma a otro y las traducciones del entorno de una app o software. Otro ejemplo de las aplicaciones de MT es la traducción de las sesiones del Parlamento de la Unión Europea. En estas sesiones se deben redactar actas en las 23 lenguas oficiales de la Unión Europea que dan pie a más de 506 combinaciones lingüísticas, ya que cada lengua puede traducirse a las otras 22 [5].

4.2.2. Clasificación de los Sistemas de MT

Los sistemas de MT se pueden clasificar en dos grupos: MT basada en reglas lingüísticas y MT basada en *corpus* lingüísticos. MT basada en *corpus* lingüísticos a su vez se dividen en MT por memoria y SMT. El esquema de clasificaciones se muestra en la figura 4.4.

[5] Página oficial del Parlamento Europeo (Oficina de Barcelona) url: http://www.europarlbarcelona.eu/es/parlament_europeu/multilinguisme/quantes_llengues.html

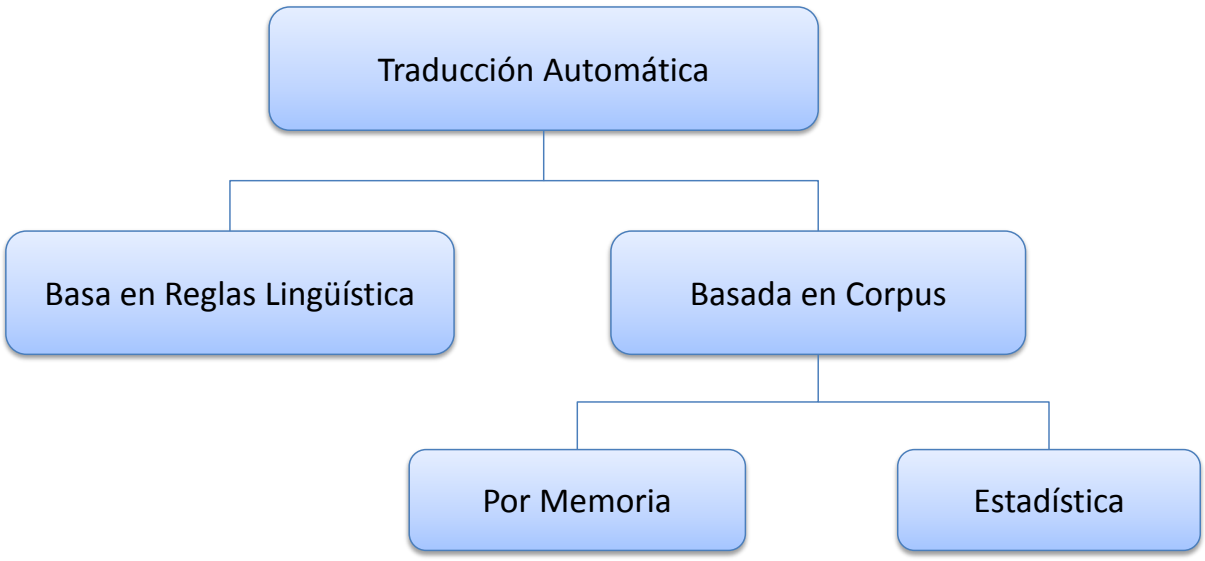


Figura 4.4: Esquema acerca de la clasificación de la traducción automática

Traducción Automática Basada en Reglas Lingüísticas

La MT basada en reglas lingüísticas consiste en una arquitectura clásica de la MT. Este enfoque, comúnmente se visualiza con el **Triángulo de Vauquois** de la figura 4.5. En el triángulo se distinguen tres enfoques principales: los enfoques directos, los de transferencia y los de interlingua. Esta pirámide se basa en las diferencias de "longitudes relativas" de los tres componentes de la traducción: análisis, transferencia y síntesis [27].

En la traducción directa, base de la pirámide de la figura 4.5, consiste en realizar una traducción palabra a palabra desde la lengua origen a la lengua destino. Esta traducción se apoya en diccionarios bilingües. En el intermedio de la pirámide de la figura 4.5, se ubica el enfoque de transferencia, que consiste en realizar un análisis gramatical desde el área sintáctica 4.2.6 y algunas veces desde la semántica 4.2.6. Finalmente, en la cima de la pirámide de la figura 4.5, se encuentra la interlengua, que consiste en un análisis semántico profundo [27].

Traducción Automática Basada en *Corpus*

La MT basada en *corpus*, consiste en que los modelos de traducción se obtienen a partir del análisis de ejemplos de un *corpus* paralelo 4.2.4. La MT basada en *corpus* a su vez se divide en MT por memoria y traducción automática estadística (*Statistical Machine Translation* - SMT).

La MT por memoria consiste en una traducción por analogía, es decir, resuelve un problema basándose en la solución de problemas similares existentes en el *corpus*. Y la SMT consiste en la generación de traducciones en base a modelos estadísticos y de teoría de la

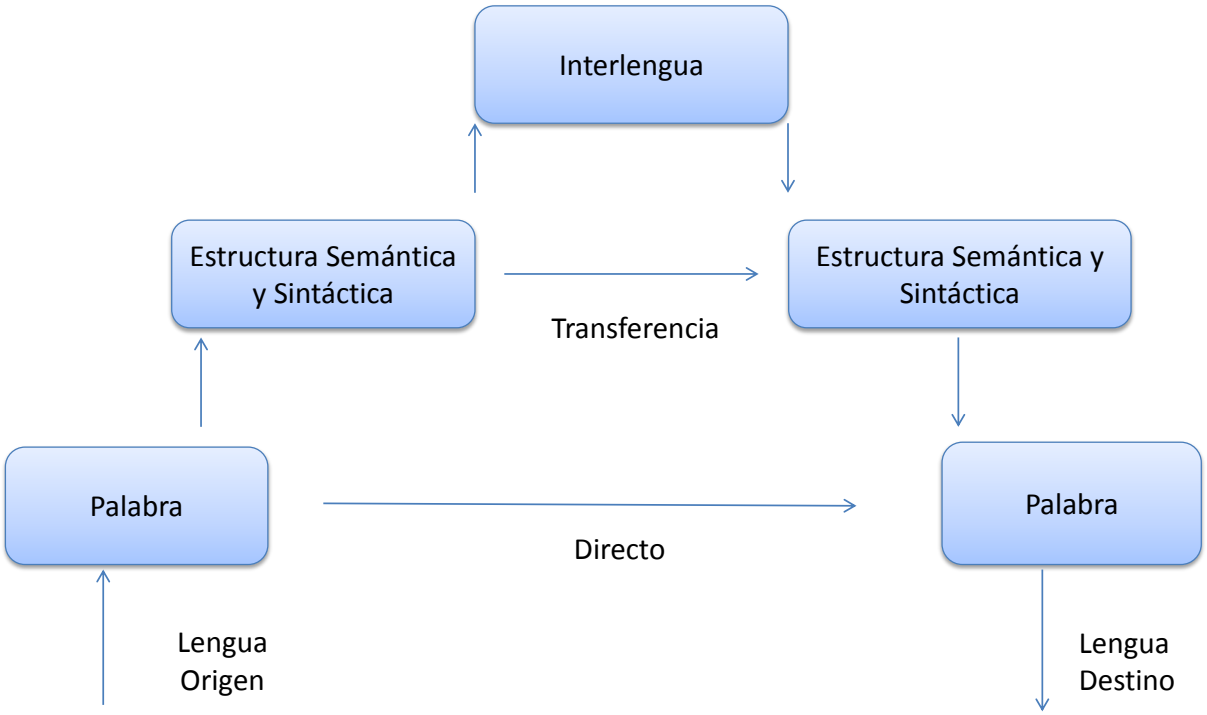


Figura 4.5: Triangulo de Vauquois

información [27]. En esta tesis, nos centramos en la SMT.

4.2.3. Traducción Automática Estadística

Las primeras ideas de traducción automática estadística (*Statistical Machine Translation* - SMT) fueron introducidas por Warren Weawer en 1949. Weawer fue un biólogo e informatólogo estadounidense que junto con Claude Shannon consolidaron la teoría de la información [28]. Entre las décadas de los cincuenta y noventa, el auge de la traducción disminuyó. Pero al aumentar la capacidad de procesamiento y almacenamiento de las computadoras y el aumento de los *corpus* paralelos disponibles para entrenar los sistemas, la MT recibió un nuevo impulso [29].

El retorno a escena de la MT se produce en 1991 con el sistema *Candide*. Este sistema fue desarrollado por un grupo de investigadores del Thomas J. Watson Center de IBM en Nueva York [9]. El sistema *Candide* fue un intento de traducción entre lenguas, donde se probó únicamente técnicas estocásticas. El sistema lo entrenaron con el *corpus Hansard* de las Actas del Parlamento Canadiense con una cantidad de aproximadamente tres millones de oraciones en inglés y francés. En el desarrollo del sistema se alinearon oraciones, grupos de palabras y palabras sueltas y después calcularon las probabilidades de que una palabra de una oración en una lengua correspondiera con otras palabras en la traducción. Los resultados

fueron muy prometedores, ya que casi la mitad de las oraciones traducidas eran exactamente como las contenidas en el texto original o tenían el mismo sentido aunque con palabras distintas. El sistema no se llegó a comercializar, pero supuso un hito histórico que dio un giro a las investigaciones. Desde 2006, la SMT es la rama de la MT más estudiada.

La SMT tiene ventajas con respecto a la MT basada en reglas, tales como, un mejor uso de los recursos, una mayor naturalidad de las traducciones y los sistemas de entrenamiento generados son fácilmente adaptables a otro par de lenguas. El principal inconveniente de la SMT es la dependencia a un *corpus* paralelo [9]. Esta última, no es tal como una desventaja, debido a que en la red existen diferentes *corpus* paralelos disponibles para los usuarios, se convierte en una desventaja precisamente en el momento en que se requiere un *corpus* en un dominio muy específico.

La SMT consiste en calcular la probabilidad $p(d|o)$ de que una cadena d de la lengua destino sea la traducción de una cadena o en la lengua origen. Esta probabilidad se calcula aplicando el Teorema de Bayes expresado en la ecuación (4.1)

$$p(d|o)\ \alpha\ p(o|d) \times p(d) \tag{4.1}$$

Donde $p(o|d)$ es la probabilidad de que la cadena origen sea la traducción de la cadena destino (modelo de traducción) y $p(d)$ es la probabilidad de ver aquella cadena destino (modelo de lenguaje). Matemáticamente, la mejor traducción se consigue escogiendo aquella que dé la probabilidad más alta, como se muestra en la (4.2)

$$\arg\max p(d|o) = \arg\max p(o|d) \times p(d) \tag{4.2}$$

La SMT se puede clasificar en tres maneras: traducción basada en palabras, traducción basada en frases y traducción basada en transductores de estados finitos. Para esta tesis, se implementó traducción basada en frases.

La SMT basada en frases, resuelve la limitación de la SMT basada en palabras. El alineamiento entre la lengua destino y la lengua origen del *corpus* paralelo no necesariamente debe contener la misma longitud. Las frases no son de carácter lingüístico, son frases encontradas en el *corpus* utilizando métodos estadísticos. A estas frases se le suelen llamar comúnmente subfrases.

Un proceso de SMT basada en frases o subfrases consta de un modelo de traducción (TM), un modelo de lenguaje (LM) y un decodificador. El modelo de traducción se obtiene a partir del alineamiento entre las frases del *corpus* paralelo y la extracción de subsecuencias de palabras. El modelo de lenguaje se obtiene por un entrenamiento con la lengua destino. Estos modelos los utiliza un decodificador para generar la traducción. Finalmente se emplearon métricas para la evaluación de la calidad del sistema de traducción. El proceso completo de

traducción y evaluación se ejemplifica en la figura 4.6.

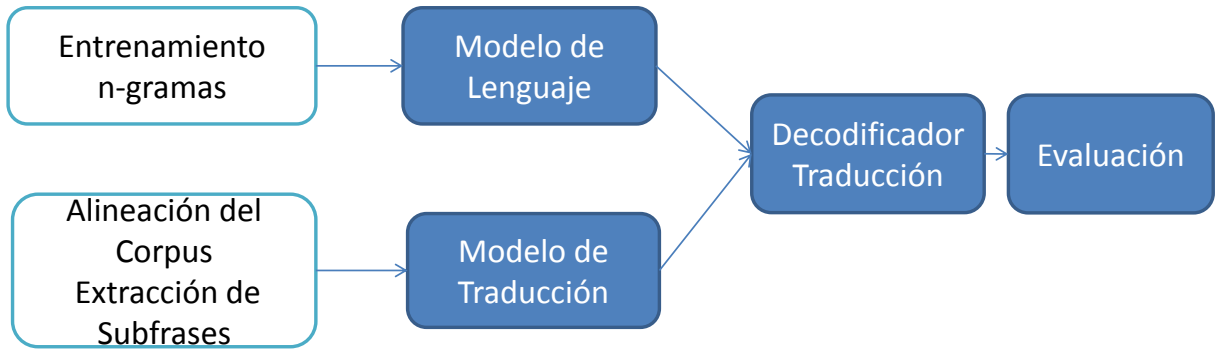


Figura 4.6: Esquema de traducción basada en frases o subfrases

El costo de la probabilidad que se asigna a una traducción es el producto de los pesos de probabilidad de los 2 modelos: modelo de lenguaje, modelo de traducción. $p(e|f)$ es la probabilidad de traducción y matemáticamente se define como:

$$p(e|f) = TM^{peso_{TM}} \times LM^{peso_{LM}} \tag{4.3}$$

Modelo de Lenguaje

El modelo de lenguaje (LM) se encarga de medir que tan probable es una secuencia de palabras, es decir, selecciona entre una lista de probables traducciones la frase con mayor ocurrencia [30]. Además, el LM toma decisiones acerca del orden correcto del vocablo y selecciona el léxico con múltiples significados o traducciones [25]. Por ejemplo:

$$P_{lm}(\text{Yo me ejercito en la casa}) > P_{lm}(\text{Ejercito yo me en la casa}) \tag{4.4}$$

$$P(\text{Fuertes tormentas en el oeste del país}) > P(\text{Musculosas tormentas en el oeste del país}) \tag{4.5}$$

En el ejemplo (4.4) el modelo de lenguaje asigna una mayor probabilidad a la primera frase, debido a que es la traducción más posible. Para el ejemplo (4.5), la palabra en español "fuerte" tiene varios sinónimos tales como musculoso, vigoroso, grueso o corpulento; el modelo de lenguaje selecciona la palabra más ajustada al contexto siendo en este caso "fuerte".

De los modelos de lenguaje probabilísticos más utilizado para la SMT, es el modelo basado en **n-gramas**. El LM de tipo **n-gramas** pretende calcular la probabilidad de una palabra basada en las palabras anteriores, es decir, se basa en la historia para predecir la palabra

siguiente. El grado del modelo (n), indica que el modelo se basa en el contexto de las n-1 palabras anteriores. Si el modelo es de segundo orden (n=2) se denomina **bigrama**, si es de tercer orden (n=3) se denomina **trigrama**.

Para hallar la probabilidad, se usa la regla de la cadena expresada en la ecuación (4.6) para el caso de una frase de 4 palabras en total.

$$P(a,b,c,d) \;=\; P(a) \times P(b|a) \times P(c|b,a) \times P(d|c,b,a) \tag{4.6}$$

Cada uno de los factores de la ecuación (4.6) se halla al contar la frecuencia en el cuerpo de entrenamiento de las secuencias de palabras, como se muestra en la ecuación (4.7).

$$\begin{aligned} P(b|a) &= \frac{Frecuencia\ de\ las\ subfrases\ (a,b)}{Frecuencia\ de\ lassubfrases(a,x)} \\ P(c|b,a) &= \frac{Frecuencia\ de\ las\ subfrases\ (a,b,c)}{Frecuencia\ de\ las\ subfrases\ (a,b,x)} \\ P(d|c,b,a) &= \frac{Frecuencia\ de\ las\ subfrases\ (a,b,c,d)}{Frecuencia\ de\ las\ subfrases\ (a,b,c,x)} \end{aligned} \tag{4.7}$$

Donde x es cualquier palabra.

Pero al aplicar la regla de la cadena a textos muy amplios, el conteo de las frecuencias será demasiado extenso, así que se recurre a aplicar la **Cadena de Márkov**. La Cadena de Márkov consiste en una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior, este postulado se expresa en la ecuación (4.8).

$$P(w_1, w_2, ..., w_n) \;\approx\; \prod_{1}^{i} P(w_i|w_{i-k}, ..., w_{i-1}) \tag{4.8}$$

Donde:

n : Número total de palabras
i : Es un contador que itera por cada palabra de la frase
k : Número de gramas (k-gramas)

A continuación se mostrará un ejemplo con la secuencia “Yo viajo a Cartagena cada año” con un orden de 2-gramas o bigramas:

$$P(Yo, viajo, a, Cartagena, cada, \text{año}) = P(Yo| < s >) \times P(viajo|yo) \times P(a|viajo) \times P(Cartagena|a) \times P(cada|Cartagena) \times P(\text{año}|cada) \times P(< /s > |\text{año})$$

Y un orden de 3-gramas o trigramas:

$$P(Yo, viajo, a, Cartagena, cada, \text{año}) = P(Yo| < s >, < s >) \times P(viajo| < s >, yo) \times P(a|yo, viajo) \times P(Cartagena|viajo, a) \times P(cada|a, Cartagena) \times P(\text{año}|Cartagena, cada) \times P(< /s > |cada, \text{año})$$

Cabe resaltar que los inicios de linea se indican con el marcador $< s >$ y finales de linea con $< /s >$. Para *corpus* mucho más extensos, las probabilidades son más pequeñas y se corre el riesgo de que se produzca *underflow*, por lo cual es siempre recomendable trabajar en el espacio logarítmico.

Debido a que la probabilidad es una serie de productos, lo cual, si una de las probabilidades de los n-gramas es cero, se tendrá problemas para realizar el cálculo. Por tal motivo, existen técnicas de suavizado para evitar probabilidades cero producidas por n-gramas no vistos, técnicas tales como el método de Aproximación o Descuento de *Laplace*, métodos de interpolación, método de *Back-Off*, método de *Witten-Bell*, *Add-One*, *Add-*α, *Good–Turing* o *Kneser–Ney* [25].

El *Add-One* consiste en agregar un uno (1) a cualquier evento. Debido a la distorsionada distribución Zipfian de palabras y n-gramas en un lenguaje natural, esto tiende a sobre estimar eventos no vistos. La sobrestimación de eventos no vistos se puede aliviar con el método *Add-*α que consiste en no agregar un uno (1) sino un valor pequeño denominado α ($\alpha = 0.00017$) a los eventos no vistos. Otra forma de suavizado es el de *Good–Turing* que no requiere datos de validación retenidos. Este método consiste en que dado un conjunto de observaciones pasadas se puede predecir la probabilidad de un evento no visto. Otra forma de suavizado son los métodos de interpolación que combinan modelos de n-gramas de orden superior con modelos de n-gramas de orden inferior. Otro método es el método de *Back-Off* que estima la probabilidad de una palabra dada su historia en el n-grama. Este modelo realiza la estimación “retrocediendo” a modelos con historiales más pequeños bajo ciertas condiciones. El método de *Witten-Bell* tiene en cuenta la diversidad de palabras pronosticadas para un historial determinado. El suavizado de *Kneser-Ney* también considera la diversidad de historias de una determinada palabra pronosticada para la estimación de modelos de reducción de orden inferior. El *Kneser-Ney* modificado usa el descuento absoluto, el cual resta un número fijo D

de cada conteo observado (cálculo especial de conteos) [25].

Existe una medida utilizada en la teoría de la información que cuantifica la calidad del modelo del lenguaje. Esta medida se denomina **perplejidad**. La perplejidad mide la incertidumbre de las predicciones y es una forma de cuantificar qué tan bien el modelo puede predecir la siguiente palabra. Un buen modelo es aquel que asigne una alta probabilidad a la palabra que efectivamente ocurre. A menor perplejidad, mejor es el modelo [30]. La perplejidad está basada en la **entropía cruzada**. La entropía cruzada se define como se muestra en la ecuación (5.3)

$$H_{P_{LM}} = \frac{-1}{n} \log P_{LM}(w_1, w_2, ..., w_n) = \frac{-1}{n} \sum_{i=1}^{n} \log P_{LM}(w_i | w_1, ..., w_{i-1}) \tag{4.9}$$

Así que la perplejidad, es finalmente:

$$PP = 2^{H_{P_{LM}}} \tag{4.10}$$

Modelo de Traducción

El modelo de traducción (*translation model* - TM) define que frases son la traducción de otras frases con su respectiva probabilidad. Estas probabilidades se muestran mediante las **tablas de traducción**.

El modelo de traducción (TM) se construye con una serie de pasos sencillos. El primero es alinear el *corpus* paralelo en ambas direcciones, es decir, cada palabra en la lengua origen está alineada con una palabra en la lengua destino. En algunos casos, es posible que hayan palabras que no estén alineadas con ninguna otra como se muestra en la figura 4.11

	cuánto	cuesta	un	certificado	de	notas
CERTIFICADO				X		
NOTA						X
PAGAR		X				
CUÁNTO	X					

Cuadro 4.11: Ejemplo de una frase en glosa alineada con una frase en español

El siguiente paso es realizar la extracción de subfrases consistentes, esto quiere decir, que todos los puntos de alineamiento para filas y columnas a los que toque la caja de color verde deben estar dentro de ella, no fuera. En la tabla 4.12 se muestra un ejemplo de extracción de una subfrase consistente y en la tabla 4.13 de extracción inconsistente de un alineamiento de inglés a español.

	cuánto	cuesta	un	certificado	de	notas
CERTIFICADO				X		
NOTA						X
PAGAR		X				
CUÁNTO	X					

Cuadro 4.12: Ejemplo de una extracción de una subfrase consistente en la alineación de un *corpus*

	cuánto	cuesta	un	certificado	de	notas
CERTIFICADO				X		
NOTA						X
PAGAR		X				
CUÁNTO	X					

Cuadro 4.13: Ejemplo de una extracción de una subfrase inconsistente en la alineación de un *corpus*

El último paso consiste en generar las probabilidades de traducción léxica más probable a cada subfrases extraídas en el paso anterior. La herramienta o *toolkit* extrae las tablas de probabilidad de traducción como se muestran en la figura 7.1:

$$House \mid Casa \mid 0{,}3 \tag{4.11}$$

En la ecuación 7.1, la tabla expresa que la probabilidad de la traducción en español de la palabra inglesa "House" sea "Casa" es del 0,3; escrito matemáticamente es $P(Casa|House) = 0{,}3$.

Las tablas de traducción son la principal fuente de conocimiento para el decodificador en la MT. El decodificador consulta estas tablas para descubrir cómo traducir la entrada en un idioma a la salida en otro idioma. Al ser un modelo de traducción de frases, las tablas de traducción no solo contienen entradas de palabras únicas, sino entradas de varias palabras [31]. Por ejemplo:

$$The\ house\ is\ |La\ casa\ es\ \mid 0{,}8$$

Generar tablas de traducción con entradas en frases y no solo en palabras, ayuda a resolver ambigüedades de la traducción. Además, si se tiene un gran texto de entrenamiento, el sistema puede aprender frases más largas, incluso hasta memorizar la traducción de oraciones

completas.

Decodificador

Como se muestra en la figura 4.6, el decodificador es el encargado de la traducción. Para este procedimiento, rigurosamente se debería realizar una búsqueda exhaustiva pasando por todas las cadenas de la lengua destino. Sin embargo, el decodificador realiza esta búsqueda de manera más eficaz al limitar el espacio de búsqueda y al mismo tiempo mantener una calidad aceptable. La función del decodificador es encontrar la mejor traducción con base en las tareas realizadas por los 2 modelos.

Una vez obtenidas las subfrases consistentes de palabras de una frase se genera un espacio de búsqueda. Finalmente se realiza la selección del camino más óptimo. Por ejemplo: Siendo una frase P1 P2 P3 P4 P5 en lengua origen, existen las siguientes posibilidades de subsecuencias:

P1 P2 | P3 P4 P5
P1 | P2 P3 | P4 P5
P1 | P2 | P3 | P4 P5

Cada posible subfrase es puesta en un nodo de color azul como se muestra en la 4.7. Posteriormente se identifican los posibles caminos, para el caso del ejemplo, el camino más optimo es el de color naranja. Notese que en la figura 4.7 existen caminos demarcados con una X debido a que son caminos no posibles. Suponiendo que se empieza por P1 P2, en el siguiente paso, el nodo P1 P2 ya no se alcanzará, por estar ya traducidas estas palabras. Igualmente, si se empieza a traducir por P3, no se podrá seguir traduciendo el segmento P2 P3. Lo mismo se repite hasta que todas las palabras hayan sido traducidas, y se obtiene finalmente el camino óptimo. Finalmente, se realiza la búsqueda hacia atrás (*backtracking*), obteniendo así la secuencia de signos deseada.

4.2.4. *Corpus*

Para realizar la traducción, se requiere una recopilación de texto o frases que tengan un significado, a esto se le denomina ***corpus***. Un *corpus* consiste en la base de datos con la cual el sistema de traducción es entrenado. Para el caso específico de traducción, se requiere un ***corpus* paralelo**, que básicamente son una colección de textos emparejados o alineados con una traducción en otro idioma. Ejemplo:

$$LA\ CASA\ ES\ BLANCA\ =\ THE\ HOUSE\ IS\ WHITE$$

Un sistema de traducción entrenado con un *corpus* que tenga como tema común un punto

de información de un puesto de salud no será igual de eficiente al ser utilizado en una sesión del Senado de la República. La delimitación del tema o contexto para un *corpus* se denomina **dominio**.

Restringir el dominio simplifica la traducción, aunque elegir un dominio no significa que el *corpus* sea escaso en frases o textos. Ampliar el *corpus* permite aumentar la flexibilidad del sistema entrenado y esto se puede realizar añadiendo variantes de cada frase de manera que la traducción sea la misma. En el siguiente ejemplo se muestra una posible forma de ampliar un *corpus*:

$$LA\ CASA\ ES\ BLANCA\ =\ THE\ HOUSE\ IS\ WHITE$$
$$LA\ CASA\ ES\ DE\ COLOR\ BLANCA\ =\ THE\ HOUSE\ IS\ WHITE$$
$$LA\ VIVIENDA\ ES\ BLANCA\ =\ THE\ HOUSE\ IS\ WHITE$$

Para que un *corpus* paralelo sea verdaderamente útil, debe estar alineado. El alineamiento consiste en poner las oraciones traducidas correspondiente a cada una de ellas. Esto se debe a que no siempre, las traducciones se dan palabra a palabra o frase a frase. Un ejemplo de alineación de *corpus* se muestra en la figura .

Los textos o *corpus* suelen ser procesados antes de entrenar el sistema, esto con el objetivo de aprovechar al máximo la información que contiene. Los procesos más utilizados para procesar el *corpus* son:

Tokenización: Consiste en separar cada texto en secuencias de palabras incluyendo los signos de puntuación o caracteres. Cada secuencia individual o palabra se denomina **token**. Por ejemplo se tiene el siguiente *corpus*:

¿LA CASA ES BLANCA?; NO ME GUSTA EL BLANCO.

Después de un proceso de tokenización el resultado obtenido es:

“¿” “LA” “CASA” “ES” “BLANCA” “?” “;” “NO” “ME” “GUSTA” “EL” “BLANCO” “.”

Trucase: Consiste en convertir la palabra en mayúscula o minúscula a la forma más probable. Para esto, lo primero que se debe generar es la lista de vocabulario con su respectiva frecuencia. Posteriormente, se sustituyen las palabras. Por ejemplo:

Se cuenta con el siguiente *corpus*:

¿La iglesia está ubicada dónde nos encontramos aquel día?
¿Dónde se encuentra la iglesia?
¿En dónde está la iglesia?
¿Por dónde llego a la iglesia?

Al generar el listado de vocabulario se tiene por resultado:

La (1/4) la (3/4)
iglesia (4)
está (2/2)
ubicada
dónde (3/4) Dónde (1/4)
nos
encontramos
aquel
día
se
encuentra
En
Por
llego
a

Así que, al realizar el proceso de trucase el *corpus* quedaría:

¿**la** iglesia está ubicada dónde nos encontramos aquel día?
¿**dónde** se encuentra la iglesia?
¿En dónde está la iglesia?
¿Por dónde llego a la iglesia?

Limpieza: Todos los caracteres como signos de puntuación, signos de pregunta, signos de exclamación, lineas vacías y caracteres de espacio redundantes son removidos. Para el caso de los *corpus* que son transcripciones orales, se les debe limpiar las representaciones escritas de las pausas, cambios de locutor, risas, toses, cambios de turno de palabra, intervenciones simultáneas de varios locutores, dudas, palabras truncadas, repeticiones entre otros.

4.2.5. Problemas con los *corpus* paralelos

A continuación se describen los dos principales problemas encontrados cuando se emplea un *corpus* [9].

Ruido

Los *corpus* generados de manera automática, poseen una característica particular y es la cantidad de ruido que llevan consigo. El ruido se debe a que usualmente los *corpus* se generan a partir de documentos disponibles en la red. Los documentos de la red se traducen de un idioma a otro mediante herramientas automáticas provocando posibles alineamientos no correctos. Además, dentro del documento existen frases en un tercer idioma, líneas en blanco, dobles espacio, formatos no compatibles, entre otros. Este tipo de errores en el *corpus* pueden tener un impacto negativo en el entrenamiento del sistema, así que para ello se propone la sección de preparación del *corpus* en la subsección 4.2.4

Dependencia del dominio

Los *corpus* suelen ser grandes extensiones de texto, lo que significa un amplio uso de memoria e implican un sistema de traducción lento, lo cual, en ninguno de los dos casos es deseado. Para resolver esta cuestión, el *corpus* se puede restringir a un domino especifico. Esto consiste en adaptar las frases del *corpus* a un solo contexto. ¿ Pero qué pasa si el sistema de traducción se utiliza en un ámbito diferente al inicial?. Este es uno de los principales problemas de los sistemas de traducción basado en *corpus*, debido a que su utilidad es baja en el momento en que se cambia de contexto. Para ello, se proponen dos soluciones. La primera solución es crear un nuevo *corpus* incluyendo las frases del dominio en cuestión para realizar un reentrenamiento. La segunda es la adaptación de dominio. La primera opción resulta ser un poco más costosa que la segunda. Así que actualmente, existen técnicas de adaptación de dominios lo cual resuelve de manera eficaz dicho problema.

4.2.6. Problemas de la traducción

Existen características de la lengua natural que dificultan el proceso de traducción, algunas de ellas son:

Morfología: La estructura de la palabra cambian según el pronombre utilizado, si es plural o singular, si es presente, pasado o futuro. Por ejemplo para el caso en español es "Nosotros comemos" vs "él come" y para el caso en inglés es "*He lives*" vs "*You live*".

Sintaxis: Existen reglas que determinan el orden de las palabras y la relación entre las

palabras. Por ejemplo: “Me devolvió el bolígrafo gastado”; esta oración tiene dos posibles significados. El primero es que de mucho bolígrafos, justo le prestó el bolígrafo que está gastado. El segundo posible significado es que le prestó un bolígrafo y se lo devolvió gastado. La relación de las palabras “bolígrafo” y “gastado” en el primer caso es una cualidad del sujeto y para el segundo caso es un estado del sujeto.

Semántica: Una palabra puede tener varios significados (polisemia) o dos palabras pueden coincidir en la escritura pero tener significados diferentes (homonimia). La polisemia y la homonimia hacen que varié la traducción debido a la dependencia del contexto. Por ejemplo: “cura” puede significar “sacerdote” o “remedio médico”. Para la palabra “cura” ser traducida al inglés, sería *priest* (en el caso de “sacerdote”) y *treatment* (en el caso de “remedio médico”). Las palabras polisémicas no deben confundirse con las homónimas, ya que las homónimas tienen un origen etimológico distinto y las polisémicas tienen el mismo origen etimológico.

Pragmática: Estudia el modo en que el contexto afecta o interfiere en la interpretación del significado. La pragmática depende del habla y no de un sistema de signos plasmados. por ejemplo: “¡Eres un genio!” puede ser interpretada de dos manera: la primera como un halago a una persona que es muy inteligente o puede referirse a un insulto (eres demasiado tonto) si es dicho de manera sarcástica.

Fonética: Estudia la producción sonora de la lengua. La fonética dificultad la traducción debido a las palabras homófonas. Las palabras homófonas son palabras que se escriben diferente pero suenan igual y tienen un significado distinto entre ellas. Por ejemplo: “asesinar” significa quitarle la vida a alguien y “acecinar” que significa salar la carne.

Fonología: Describe el modo en que los sonidos funcionan en una lengua en particular o en las lenguas en general. La fonología estudia el sonido de cada una de las unidades de la lengua. Por ejemplo la palabra “casa” consta de cuatro fonemas: /k/, /a/, /s/, /a/.

4.3. Métricas de Evaluación

En esta sección se describe algunas de métricas de evaluación para MT. Las métricas son utilizadas para dar una estimación de que tan buena es una traducción dada por el sistema respecto a una traducción de referencia. Las métricas de evaluación mas utilizadas son:

4.3.1. BLEU

BiLingual Evaluation Understudy (BLEU) es una de las métricas de evaluación más usadas para un sistema de MT. Es una métrica rápida, económica, independiente del lenguaje y tiene una gran correlación con las evaluaciones manuales [34]. BLEU compara los n-gramas de la frase generada por el sistema de traducción con los n-gramas de la frase de referencia, contando el número de n-gramas que coinciden independientemente de la posición. La BLEU se puede calcular utilizando mas de una traducción de referencia, lo cual permite una mayor robustez a la medida frente a traducciones libres realizadas por humanos. La BLEU se calcula mediante la expresión 5.1.

$$\text{BLEU} = BP \exp\left(\sum_{n=1}^{N} W_n \log p_n\right) \tag{4.12}$$

Donde:

N : Orden de los n-gramas calculados.
BP : Es el factor que penaliza las traducciones que sean mas cortas que su frase original
P_n : Precisión
W_n : Peso uniforme de la forma $W_n = \frac{1}{N}$

El término BP y P_n de la ecuación 5.1 se puede calcular mediante la expresión:

$$BP = \exp\left(\min\left(1, 1 - \frac{L_{ref}}{L_{sys}}\right)\right) \tag{4.13}$$

$$P_n = \frac{\sum_{c \in \mathcal{C}} \sum_{n\text{-gram} \in \text{C}} \text{Count}(n\text{-gram})}{\sum_{c \in \mathcal{C}} \sum_{n\text{-gram} \in \text{C}} \text{Count}_{sys}(n\text{-gram})} \tag{4.14}$$

Donde:

- L_{ref}: Número de palabras de referencia que tiene una longitud más parecida a la frase traducida.

- L_{sys}: Número de palabras en la frase traducida.

- Count: Número de n-gramas encontrados tanto en la frase candidata de referencia C como en la frase traducida.

- Count_{sys}: Número de n-gramas encontrados únicamente en la frase traducida.

- $\mathcal{C}$: Todas las frases candidatas de traducción.

- C: Frase de referencia.

- n-gram: n-grama.

- c: Frase Traducida.

La BLEU toma valores entre 0 y 1 y, dado que es una medida de precisión, la calidad del sistema es mejor cuanto más alta sea la medida.

4.3.2. NIST

La métrica NIST es una variante de la métrica BLEU. Estas dos métricas se diferencian porque la NIST le da mayor peso a los n-gramas menos frecuentes dentro de la traducción. La NIST se calcula mediante la siguiente expresión:

$$\text{NIST} = \sum_{n=1}^{N} \left(\frac{\sum_T \text{Info}(n\text{-gram})}{\sum_{n\text{-gram} \in s_1} 1} \right) \exp \left[\beta \log_2 \left(\min \left(1, \frac{L_{sys}}{\overline{L}_{ref}} \right) \right) \right] \tag{4.15}$$

$$\text{Info}(n\text{-gram}) = \log_2 \left(\frac{\text{Count}(w_1, ..., w_{n-1})}{\text{Count}(w_1, ..., w_n)} \right) \tag{4.16}$$

Donde:

- N : Orden de los n-gramas.

- L_{sys} : Número de palabras en la frase traducida.

- $\bar{L}_{ref}$: Número medio de palabras de una traducción de referencia, promedio con todas las traducciones de referencia.

- count(.) : Número de ocurrencias para n-gramas $(w_1, ..., w_n)$ y $(w_1, ..., w_{n-1})$ en todas las traducciones de referencia.

- β : Se elige para hacer el factor BP. Es 0,5 cuando el número de palabras en la frase traducida es 2/3 del número promedio de palabras en la traducción de referencia.

- n-gram: n-grama.

- T: Todos los n-gramas ocurridos.

La NIST toma valores entre 0 y 10 y, como también es una medida de precisión, la calidad del sistema es mejor cuanto más alta sea la medida.

4.3.3. WER

Porcentaje de Palabras Erróneas o *Word Error Rate* (WER). Está métrica mide el porcentaje mínimo de palabras que hay que insertar, eliminar o sustituir en la traducción para obtener la frase de referencia. La WER se calcula mediante el conteo del número de inserciones, borrados y sustituciones de palabras cuando se compara la traducción con la referencia. Esta medida se basa en la distancia de edición o de *Levenshtein*. La expresión para calcular la WER en cada frase de salida del traductor con respecto a la frase de referencia es:

$$\text{WER}(\,\%) = \frac{S + B + I}{N} \times 100 \tag{4.17}$$

Donde:

S : Número de palabras sustituidas.
B : Número de palabras borradas.
I : Número de palabras insertadas.
D : Número de palabras de la frase de referencia.

La métrica WER tiene como desventa que presenta una dependencia con las frases de referencia. Existe un número casi ilimitado de traducciones correctas para una misma frase, y sin embargo, esta medida considera que sólo una es la correcta [30].

4.3.4. SER

Porcentaje de frases erróneas o *Sentence Error Rate* (SER). La SER es el porcentaje de frases en donde las traducciones coinciden exactamente con las frases de referencia. Se calcula igualmente con la ecuación 5.2. La SER posee la misma desventaja de la métrica WER [9].

4.3.5. mSER / mWER

Multireference Sentence Error Rate - mSER y *Multireference Word Error Rate* - mWER. La mSER y la mWER son métricas similares a la SER y WER respectivamente, con la breve diferencia de que esta considera varias referencias para cada frase a traducir. Es decir, para cada frase se calcula la distancia de editado con las distintas referencias y finalmente quedando con la más pequeña [35].

A continuación se muestra un resumen de las características de evaluación descritos anteriormente.

Sigla	Aplica sobre	# de Ref	Rango	Descripción
BLEU	Frase	Varias	[0 -1]	Cuenta el número de n-gramas que coinciden del texto traducido con el texto de referencia independientemente de la posición.
NIST	Frase	Varias	[0-10]	Es igual que la BLEU con la diferencia de que esta le da mayor peso a los n-gramas menos frecuentes dentro de la traducción.
WER	Palabra	Una	[0-100]	Porcentaje (%) de palabras que hay que insertar, eliminar o sustituir para obtener la referencia
SER	Frase	Una	[0-100]	Porcentaje (%) de frases diferentes a la referencia
mWER	Palabra	Varias	[0-100]	Es igual que la WER pero con varias frases de referencia
mSER	Frase	Varias	[0-100]	Es igual que la SER pero con varias frases de referencia
PER	Palabra	Una	[0-100]	Es igual que la WER pero con cualquier posible orden de palabras en la referencia

Cuadro 4.14: Resumen de las principales características de las métricas de evaluación

4.4. Herramientas: Moses

Moses[6] es una herramienta bajo la licencia pública general reducida de GNU. El sistema Moses se ejecuta en Linux. Es posible ejecutarlo desde Windows pero utilizando Cygwin. Cygwin emula la ejecución de Moses en Windows. Moses es un *toolkit* de código abierto especializado para realizar sistemas de traducción con enfoque estadístico, es decir, produce una oración destino desde una oración fuente con alta probabilidad [31]. Esta herramienta es un sucesor de un proyecto similar creado por Philipp Koehn denominado *Pharaoh* como parte de su trabajo de doctorado. En 2006 Moses pasó a estar subvencionado por la Unión Europea dentro del proyecto EuroMatrix [7] (más tarde EuroMatrixPlus). Finalmente, en 2012, la Comisión Europea creó el proyecto MosesCore con el objetivo de llenar los intereses

[6]Página oficial de Moses url: http://www.statmt.org/moses/
[7]Página oficial de EuroMatrix url: http://euromatrix.net

del mundo académico y comercial en la MT de código abierto. Philipp Koehn se encuentra actualmente al cargo del mantenimiento de Moses [32].

Moses ofrece realizar traducciones bajo 2 tipos de modelos: MT basada en frases y MT basada en frases jerárquica, la diferencia es que la primera permite agregar etiquetas de mayor estructuración en el *corpus* y las subfrases extraidas del *corpus* no deben ser palabra secuenciales. Para efectos de este trabajo de grado usaremos las herramientas para la traducción automática basada en frases.

En la figura 4.9 se mostrará una descripción completa de los componentes de Moses[8] para realizar un sistema de traducción basada en frases:

4.4.1. GIZA++

GIZA++ es una implementación disponible gratuitamente de los modelos de IBM. GIZA++ es la herramienta requerida para generar alineamientos de los *corpus* paralelos. Es una extensión del programa GIZA. Fue desarrollada en 1999 por el equipo de SMT durante el taller de verano en el Centro de Lenguaje y Tratamiento del Habla de la Universidad Johns Hopkins [33]. Es una herramienta que se usa para alinear palabras o secuencia de palabras mediante la implementación de los modelos IBM 1-5.

4.4.2. KenLM

KenLM es la herramienta que por defecto usa Moses para generar el modelo de lenguaje. El KenlM calcula los modelos de lenguaje utilizando el modelo de n-gramas explicado en la sección 4.2.3 y utiliza el suavizado modificado de *Kneser–Ney* sin poda. KenLM es completamente seguro para subprocesos en Moses con múltiples hilos. KenLM es simultáneamente rápido y con poca memoria. Esta implementación se mencionó en enero de 2010 y la integración a Moses se anunció públicamente el 18 de octubre de 2010. Se denomina KenLM debido a las primeras 3 letras del nombre de su creador *Kenneth Heafield* y las letras iniciales de *Language Model*.

La función *lmplz* de Moses es el encargado de estimar el modelo. La sintaxis consta del orden del n-grama (-o), una cantidad de memoria para usar (-S) y una ubicación para colocar archivos temporales (-T) como se muestra en la ecuación 4.18

$$/\text{bin}/\text{lmplz -o 5 -S 80\% - T } /\text{tmp} < \text{text} > \text{text.arpa} \tag{4.18}$$

El formato del archivo del modelo de lenguaje es .arpa

[8] *Opción por defecto de Moses

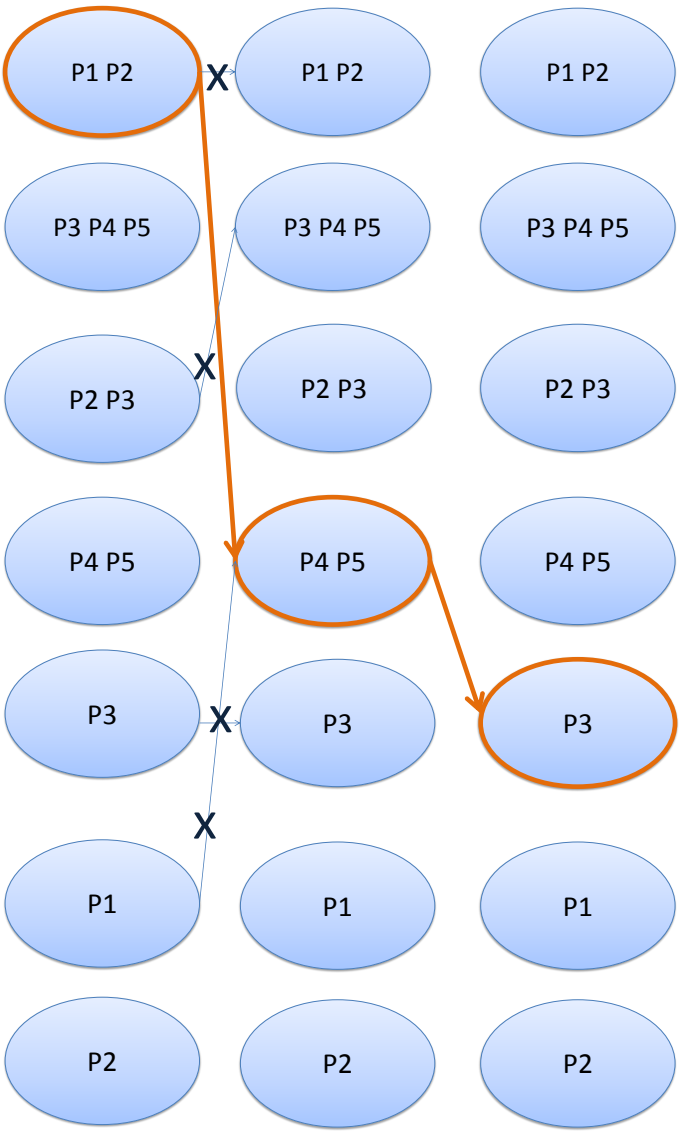


Figura 4.7: Organización de nodos y posibles caminos de traducción para la frase P1 P2 P3 P4 P5

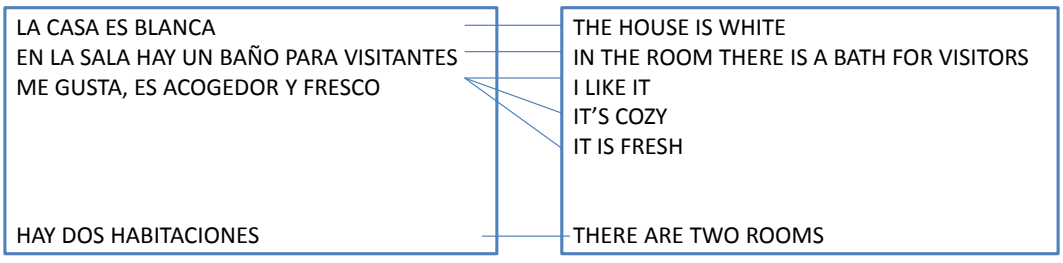


Figura 4.8: Ejemplo sobre la correcta alineación de un *Corpus* paralelo

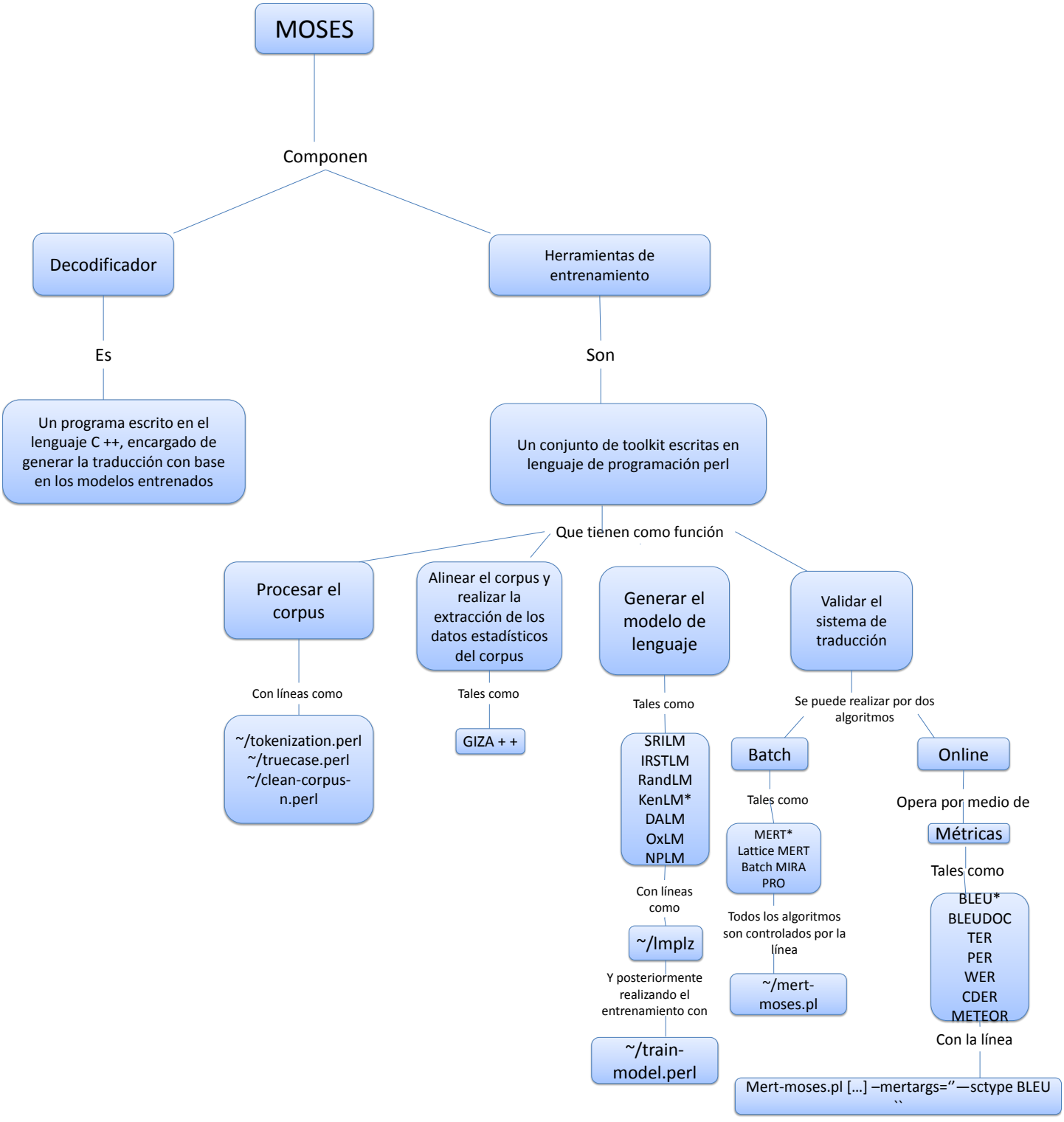


Figura 4.9: Esquema resumen de funciones de la herramienta Moses para la traducción automática basada en frases o subfrases

Capítulo 5

DESARROLLO Y RESULTADOS

En este capítulo se muestra de manera detallada todo el proceso de elaboración del sistema de traducción. La metodología propuesta para el desarrollo de esta tesis se muestra en la figura 5.1.

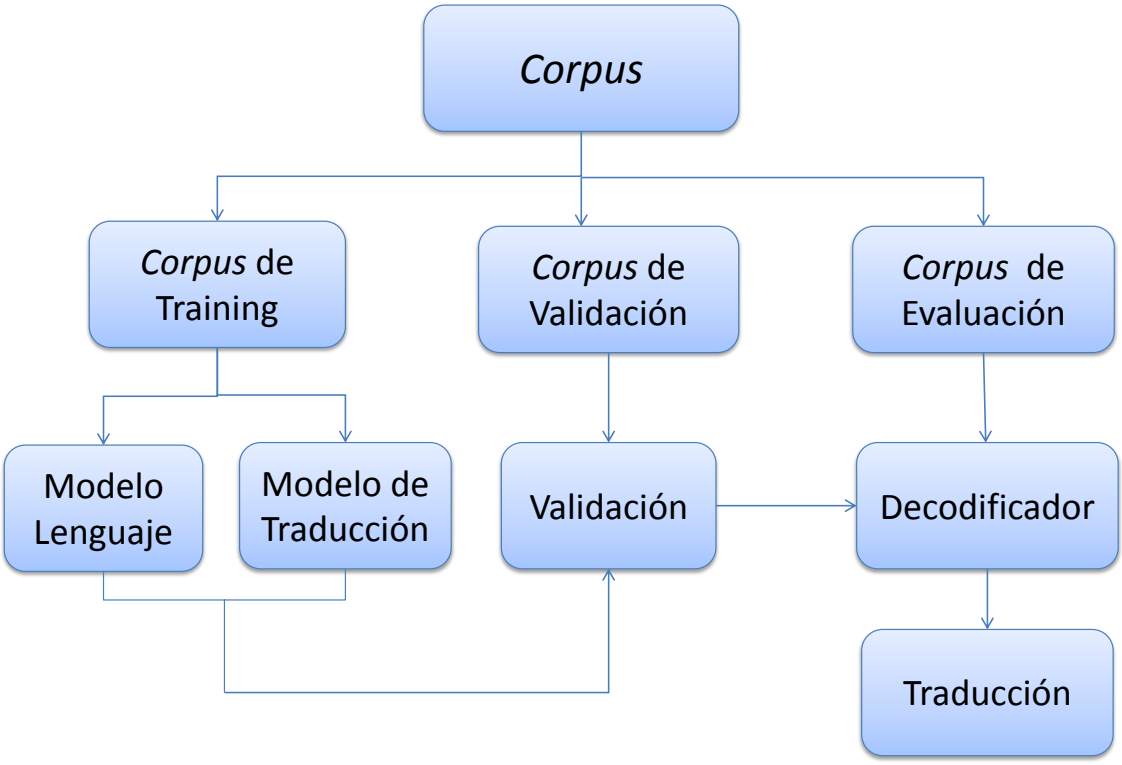


Figura 5.1: Metodología para el desarrollo de este trabajo de grado

La metodología mostrada en la figura 5.1 parte de un *corpus* divido en tres: *corpus* de entrenamiento, *corpus* de validación y *corpus* de evaluación. Con el *corpus* de entrenamiento

se realiza el modelo de lenguaje y el modelo de traducción. Una vez obtenidos los modelos se procede, junto con el *corpus* de validación, a realizar el ajuste del sistema. Finalmente, el decodificador toma el *corpus* de evaluación y genera la traducción.

En la sección 5.1 se profundiza sobre la generación y procesamiento del *corpus* paralelo, en la sección 5.2 se amplia lo perteneciente a la generación de los modelos de traducción, modelo de lenguaje. Finalmente en la sección 5.3 se aborda la validación y evaluación del sistema de traducción.

5.1. Generación y Procesamiento del *Corpus* Paralelo

Este *corpus* fue construido y ajustado con fines académicos para la realización de este trabajo de grado.

Los pasos que se siguieron para la generación del *corpus* paralelo fueron:

1. Elección del dominio y delimitación del contexto. Como se mencionó en la sección 4.2.4, elegir un tema especifico para la elaboración del *corpus* aporta a la calidad de la traducción. Para efecto de este trabajo de grado se eligió un contexto académico, específicamente el punto de información de un centro de educación superior. Si bien, el contexto es académico, no implica que el *corpus* contenga frases comunes de todo el entorno estudiantil como lo son las aulas, las cafeterías, la bibliotecas entre otros.
2. Selección de las frases en español más comunes en un punto de información académico. Se realizó mediante la consulta de las preguntas más frecuentes en los centros de educación superior en roles de estudiante, aspirante, administrativo, egresado o externo.
3. Ampliación del *corpus*. Esto se hizo añadiendo frases en español que son sinónimos de las frases ya existentes o que tienen la misma traducción en LSC. Esto es con el objetivo de robustecer el sistema entrenado. Para este paso, se incluyen sinónimos de nombres, adjetivos o verbos y se modifica la sintaxis de las oraciones.
4. Traducción de las frases del *corpus* en español a glosas por una persona experta en LSC.

En la construcción del *corpus*, inicialmente se recogieron un total de 863 frases, posteriormente se depuraron las frases ajenas al contexto y frases que generaron errores de alineación por la diferencia de longitudes (longitud de la frase en español y la longitud en la frase en glosa correspondiente), quedando finalmente 517 frases.

Las frases del *corpus* paralelo están contextualizadas dentro del dominio académico, específicamente, solicitudes que posiblemente se podrían plantear en un punto de información al

interior de una institución de educación superior. El ambiente que se pretende simular en este sistema de interacción es unidireccional (LSC a español), así que la tendencia de las frases son de carácter interrogativo, es decir, la persona sorda se acerca al punto de información, con el objetivo, en la mayoría de los casos, de formular una consulta.

En la tabla 5.1 se muestra una breve descripción del *corpus* paralelo:

Descripción	*Corpus* en Español	*Corpus* en Glosas
Número total de frases	517	517
Número de preguntas simples	165	165
Número de preguntas compuestas	323	323
Número de afirmaciones	18	18
Número de saludos y despedidas	11	11
Longitud promedio de la frase	6.26	4.11
Número de palabras que contiene la frases más larga	13	9
Número de palabras que contiene la frases más corta	1	1
Número total de tokens	3237	2128
Número total de palabras	650	390
Palabra más común	de	dónde
Número de veces que aparece la palabra más común	238	132
Número de bigramas	1660	1277
Número de trigramas	2001	1633

Cuadro 5.1: Descripción general del *corpus* paralelo empleado para este trabajo de grado

El *corpus* en español, tanto para entrenamiento, validación y evaluación, está escrito en primera persona. El *corpus* en LSC tiene por características que la glosa que representa un verbo está escrito en infinitivo, las frases omiten los pronombres personales, los objetos están en singular y se omiten los signos de puntuación. Las palabras utilizadas en el *corpus* no son propias de un solo centro educativo, este *corpus* esta escrito de manera general para ser usado por cualquier centro de educación del país.

El archivo se elaboró en un documento de texto sencillo o texto plano, el editor de texto es compatible con UTF-8, GNU/Linux, Mac OS X y Microsoft Windows.

En el desarrollo de la metodología, el *corpus* paralelo se dividió 10 veces de manera diferente como se muestra en la figura 5.2. El 90 % de las oraciones fueron destinadas para un *corpus* de entrenamiento, el 5 % de las oraciones para un *corpus* de validación o ajuste y el 5 % de las frases restantes se destinaron para el *corpus* de evaluación. La división del *corpus* paralelo se realizó de manera aleatoria mediante un código ejecutado en el software Matlab.

La limpieza del *corpus* paralelo se realizó de forma manual, debido a que el *toolkit* de Moses no reconoce algunos caracteres del español, dificultando la limpieza de este. Así que la

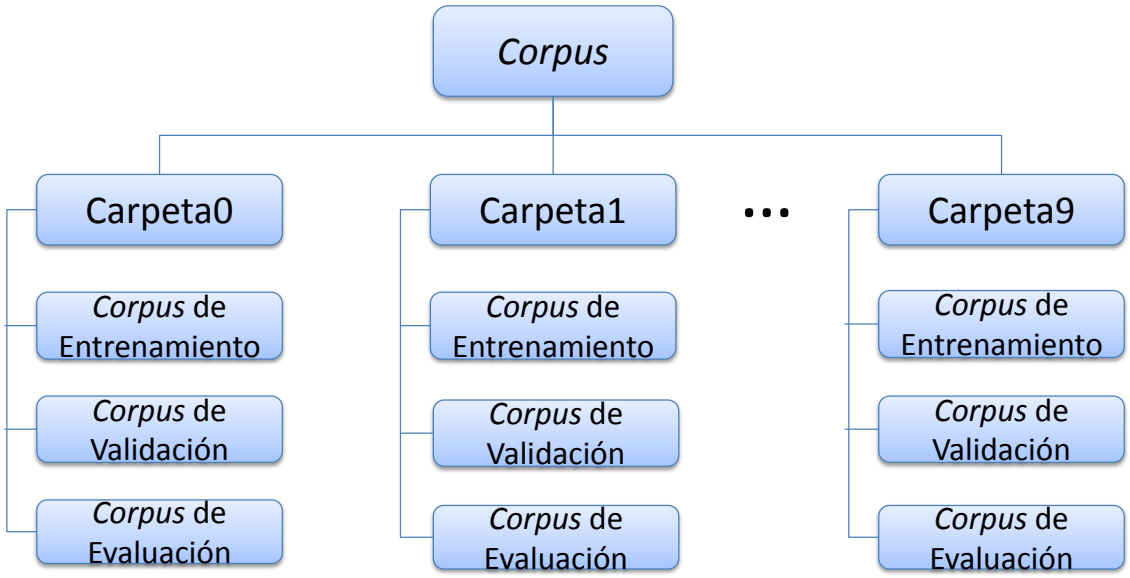


Figura 5.2: División del *corpus* en diez carpetas diferentes

edición del texto se desarrolló mediante el software Notepad++ para Unix. Se tuvo en cuenta la codificación de los finales e inicios de linea, las mayúsculas, la cantidad de espacios entre cada token y la ortografía.

5.2. Generación de los Modelos

Una vez realizado el procesamiento del *corpus*, se procede a generar los modelos de lenguaje y de traducción mediante las herramientas mencionadas en la sección 4.4.

Modelo de Lenguaje

Una vez entrenado el *corpus* de entrenamiento de la lengua destino (español) se generó un archivo .text donde se en listan el número total de unigramas, bigramas y trigramas. El archivo .text también muestra la probabilidad de cada unigramas, bigramas y trigramas existente. Para ver más detalle de las funciones y parámetros utilizados para la generación del LM ver el anexo 1 en la sección 7. En la tabla 5.2 se encuentran los resultados obtenidos para el LM de cada carpeta.

Carpeta	Número de 1-grama	Número de 2-gramas	Número de 3-gramas
Carpeta0	629	1515	1802
Carpeta1	634	1536	1832
Carpeta2	634	1536	1832
Carpeta3	632	1540	1841
Carpeta4	632	1544	1832
Carpeta5	633	1539	1829
Carpeta6	634	1531	1828
Carpeta7	625	1525	1821
Carpeta8	631	1543	1840
Carpeta9	632	1540	1841
Promedio	632	1535	1830

Cuadro 5.2: Cantidad de unigramas, bigramas y trigramas resultantes de la generación del modelo de lenguaje para cada carpeta

Para desarrollo de esta tesis, se generó con el *corpus* de entrenamiento de cada carpeta dos modelos de lenguaje: LM con bigramas y un LM con trigramas.

Evaluación del Modelo de Lenguaje

Con el objetivo de evaluar la calidad de los modelos de lenguajes generados, se calculó la perplejidad para cada uno de los modelos de lenguaje. En la tabla 5.3 se muestra los resultados de dicha evaluación.

	Perplejidad del Modelo de Lenguaje (ppl)	
Carpeta	**LM con bigramas**	**LM con trigramas**
Carpeta0	84,15967	80,99791
Carpeta1	40,52971	34,47712
Carpeta2	40,52971	34,47712
Carpeta3	29,24095	24,91227
Carpeta4	22,28311	20,78324
Carpeta5	20,85669	17,72731
Carpeta6	28,42083	24,99304
Carpeta7	28,31305	26,34238
Carpeta8	21,37116	18,32536
Carpeta9	29,24095	24,91227
Promedio	28,97	25,26

Cuadro 5.3: Cálculo de la perplejidad para los modelos de lenguaje de cada carpeta

Modelo de Traducción

En esta sección se realiza el alineamiento del *corpus* en LSC con el *corpus* en español. Seguidamente se realizó la extracción de las tablas de traducción. La herramienta Moses, de manera simultanea, genera otros modelos o funciones de puntuación de traducción de frase adicionales a las tablas de traducción los cuales son: *distortion*,*word penalty*, *phrase penalty* y *reordering model*.

El *distortion* realiza una especie de suavizado debido a que algunos pares de frases infrecuentes pueden causar problemas, es decir, si las dos frases e y f solo aparecen una vez, entonces $P(e|f) = P(f|e) = 1$ y esto a menudo sobrestima la confiabilidad de los pares de frases poco comunes. Así que, el *lexical weighting* lo que hace es descomponer las traducciones de palabras para verificar qué tan bien coinciden. El *word penalty* tiene como función proteger una producción demasiado corta o demasiado larga en la salida del sistema. El *phrase penalty*, al igual que el *word penalty*, tiene como función proteger una producción demasiado corta o demasiado larga, con la diferencia de que el *word penalty* penaliza palabras y el *phrase penalty* penaliza frases. Por último, el *reordering model* tiene como finalidad restringir o ampliar el reordenamiento a movimientos locales cortos o largos suficientes para la traducción. El *reordering model* de la herramienta Moses generalmente castiga el movimiento.

Generación del Archivo moses.ini

Una vez obtenido los modelos, se procedió a generar el archivo moses.ini. En el documento moses.ini se muestran los pesos asociados a cada modelo. Para ver más detalle de las funciones y parámetros utilizados en esta sección ver el anexo 3 en la sección 7. En las tablas 5.4 y 5.5 se muestran los pesos asignados por Moses a cada carpeta. Estos resultados muestran el peso del LM tanto para bigramas como para trigramas.

Carpeta	**Peso TM**	**Peso LM**	**Peso RM**
Carpeta0	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta1	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta2	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta3	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta4	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta5	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta6	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta7	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta8	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3
Carpeta9	0.2 0.2 0.2 0.2	0.5	0.3 0.3 0.3 0.3 0.3 0.3

Cuadro 5.4: Pesos de cada uno de los modelos determinados en el archivo moses.ini de las diez carpetas

Carpeta	Peso *word penalty*	Peso *phrase penalty*	Peso *distortion*
Carpeta0	-1	0.2	0.3
Carpeta1	-1	0.2	0.3
Carpeta2	-1	0.2	0.3
Carpeta3	-1	0.2	0.3
Carpeta4	-1	0.2	0.3
Carpeta5	-1	0.2	0.3
Carpeta6	-1	0.2	0.3
Carpeta7	-1	0.2	0.3
Carpeta8	-1	0.2	0.3
Carpeta9	-1	0.2	0.3

Cuadro 5.5: Pesos de cada uno de los modelos determinados en el archivo moses.ini de las diez carpetas

En las tablas 5.4 y 5.5 se puede notar que la herramienta Moses por defecto asigna unos pesos a los modelos, iguales para todas las carpetas. Por dicha razón, se procedió a realizar un ajuste a los pesos, este con el objetivo de encontrar mejores valores que mejoren la calidad de la traducción.

Para ver más detalle de las lineas utilizadas para generar el modelo de traducción y el archivo moses.ini dirigirse el anexo 2 de la sección 7.

5.3. Validación y Evaluación del Sistema

Debido a que Moses arroja unos pesos predeterminados a los modelos, se procedió a realizar un ajuste a dichos pesos. Para efecto de esta tesis se generaron cuatro experimentos más. Cada experimento contiene una modificación de los pesos de cada modelo. La modificación de los pesos se realizó con criterios establecidos por las autoras. Una vez ajustados los pesos de los modelos, se procede a la realizar la traducción mediante el decodificador con cada uno de los experimentos. El sistema de traducción fue evaluado con la métricas BLEU y WER. Además, cada resultado muestra un intervalo de confianza del 95 % con el objetivo de ver si es estadísticamente significativo. Dependiendo de la métrica de referencia utilizada, el intervalo se calcula mediante la ecuación 5.1 y la ecuación 5.2.

$$\pm \triangle = 1,96\sqrt{\frac{BLEU(100 - BLEU)}{n}} \tag{5.1}$$

$$\pm \triangle = 1,96\sqrt{\frac{WER(100 - WER)}{n}} \tag{5.2}$$

Donde n corresponde al número total de palabras del *corpus* de evaluación de la lengua destino (español).

En la tabla 5.6 se muestra los resultados obtenidos para las métricas de evaluación del sistema de traducción empleando un modelo de lenguaje de bigramas y en la tabla 5.7 se muestra los resultados obtenidos para el sistema de traducción utilizan un modelo de lenguaje de trigramas.

Experimento	**BLEU (%)**	**WER (%)**
moses.ini	27,29 ±6,79	9,09 ±4,54
moses1.ini	24,50 ±6,79	9,57 ±4,54
moses2.ini	25,60 ±6,79	10,31 ±4,54
moses3.ini	24,17 ±6,79 ±0,53	9,48 ±4,54
moses4.ini	25,45 ±6,79	9,55 ±4,54

Cuadro 5.6: Resultados de las métricas de evaluación BLEU y WER obtenidos para cada experimento utilizando un modelo de lenguaje de bigramas

Experimento	**BLEU (%)**	**WER (%)**
moses.ini	28,56 ±6,70	9,37 ±4,54
moses1.ini	25,34 ±6,70	9,90 ±4,54
moses2.ini	28,12 ±6,70	9,92 ±4,54
moses3.ini	23,75 ±6,70	9,61 ±4,54
moses4.ini	26,50 ±6,70	9,67 ±4,54

Cuadro 5.7: Resultados de las métricas de evaluación BLEU y WER obtenidos para cada experimento utilizando un modelo de lenguaje de trigramas

En las tablas 5.6 y 5.7, la primera columna corresponde a los diferentes pesos utilizados para la validación del sistema.El archivo moses.ini corresponde a los pesos dados por defecto desde Moses y los archivos moses1.ini, moses2.ini, moses3.ini y moses4.ini corresponden a los pesos modificados. La segunda columna de las tablas corresponde al promedio de la métrica BLEU aplicada en cada carpeta. Es decir, cada carpeta se evaluó de manera individual con cada uno de los archivos .ini y finalmente se obtuvo el promedio.

En la figura 5.3 se muestra el gráfico asociado a la evaluación del sistema de traducción con la métrica BLEU y en la figura 5.4 se muestra el gráfico asociado a la evaluación del sistema de traducción con la métrica WER.

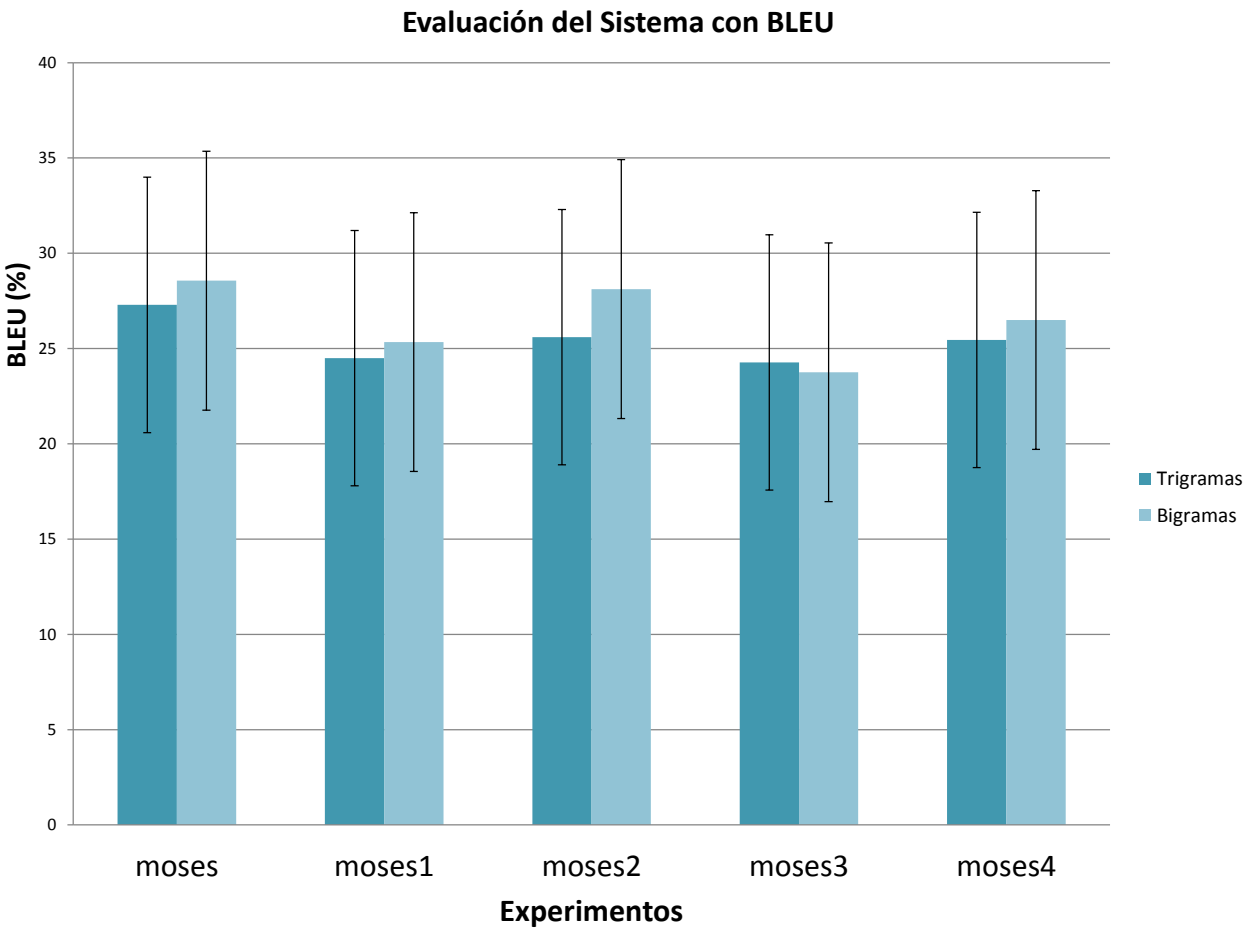


Figura 5.3: Gráfica de los resultados obtenidos para la métrica BLEU utilizando un modelo de lenguaje de trigramas y bigramas en cada experimento

Las tablas 5.6 y 5.7 muestran un resultado muy bueno para la métrica WER, diferente para el caso de la métrica BLEU. Estos resultados tienen su razón de ser. El principal motivo tiene que ver con el hecho de que, en el *corpus* paralelo generado contiene una frase en español para la traducción de una frase en glosas. Lo cual es una dificultad para la métrica BLEU debido a que esta requiere varias referencias de traducción para obtener un buen resultado. Para el caso de esta tesis solo se posee una referencia para cada frase en glosa.

Teniendo en cuenta los intervalos de confianza de la figura 5.4 se muestra una diferencia estadística no significativa entre los experimentos de los archivos .ini. Al variar los pesos de los modelos el sistema de traducción, las traducciones siguen estando dentro del intervalo de confianza.

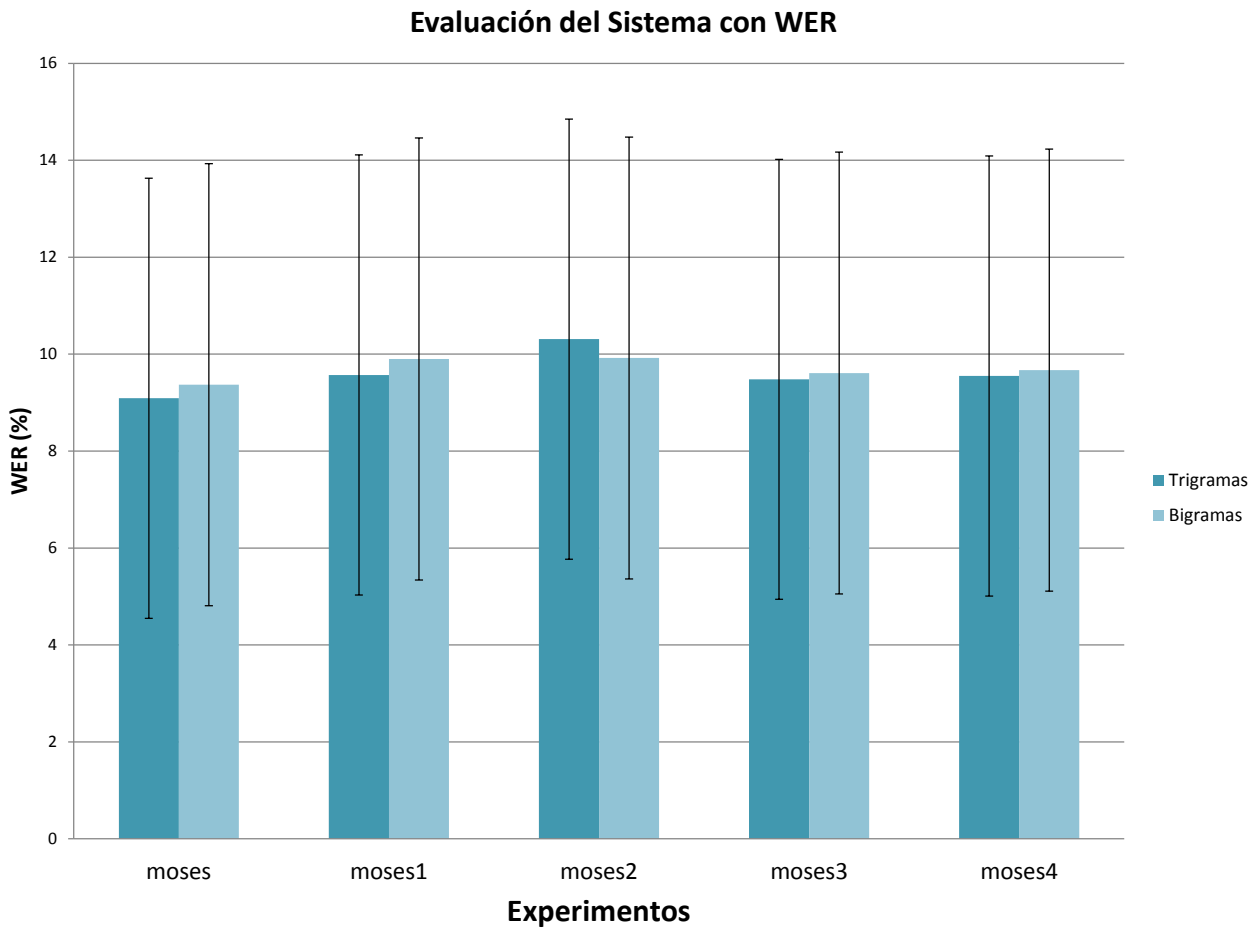


Figura 5.4: Gráfica de los resultados obtenidos para la métrica WER utilizando un modelo de lenguaje de trigramas y bigramas en cada experimento

Capítulo 6

CONCLUSIONES

En esta tesis se ha contribuido a un sistema de traducción que permite la comunicación entre personas oyentes y sordas. La elaboración de este proyecto permitirá a los sordos poder acceder con más facilidad a la información referente a un puesto de consulta de una institución de educación superior. El resultado del proyecto ha sido muy satisfactorio, a pesar de tener un *corpus* limitado y de no poder traducir cualquier contexto.

En concreto, se han alcanzado los objetivos descritos en el capítulo 3.

1. Se revisaron diferentes trabajos asociados a desarrollos de traducción automática estadística en los cuales se resume que las técnicas estocásticas dan mejor robustez a los sistemas de traducción en comparación con sistemas basado en reglas lingüísticas. Para efectos de esta tesis se usaron herramientas de procesamiento de texto y analisis estadísticos. Además de implementar una herramienta especializada en traducción automática estadística denominada Moses, a la cual, previamente se realizó la consulta bibliográfica previa para la comprensión teórica de su funcionamiento.

2. A pesar del fácil acceso a los corpus en la red para realizar traducción automática estadística, para efectos de esta memoria se construyó un *corpus* paralelo contando finalmente con 517 frases alineadas de la Lengua de Señas Colombiana al español. El *corpus* paralelo se construyó dentro de un dominio académico, específicamente un punto de información de una institución de educación superior. Si bien, delimitar el contexto del *corpus* genera una mejor traducción, existen otros factores que influyen en la calidad de la traducción, tales como son la cantidad de frases de entrenamiento, la diversidad de vocabulario y la diferencia de longitud entre las frases. Cabe resaltar que para la construcción del *corpus* fue importante contar con una persona especializada en LSC.

3. Para el procesamiento del *corpus* se recurrió a herramientas de procesamiento de texto diferentes a las brindadas por Moses. Esto se debe a que las herramientas del sistema Moses funcionan muy bien para otras lenguas como la lengua inglesa. Pero a la hora de

procesar texto en español presenta inconvenientes debido a que por ejemplo en la lengua inglesa no existen tildes ni signos de interrogación y exclamación de apertura, diferente a la lengua española. Dichos caracteres no fueron identificados y fueron pasados por alto al momento de la limpieza. Las herramientas de procesamiento de texto se deben seleccionar según el par de lenguas a trabajar.

4. El resultado con la métrica WER en promedio fue de 8,18 % ±4,54 y el resultado con la métrica BLEU en promedio fue 25,93 % ±6,70. El principal motivo del resultado de la BLEU tiene que ver con el hecho de que esta métrica requiere varias referencias de traducción para obtener un buen resultado. Para el caso de esta tesis solo se posee una referencia para cada frase en glosa. Otras posibles causas de errores de traducción del sistema están relacionadas con la distinta variabilidad de las palabras en español, el distinto orden de la oraciones en español y que las versiones de las herramientas utilizadas no son las más recientes.

Capítulo 7

TRABAJOS FUTUROS

En este capítulo se describen los principales trabajos a futuro que podrían seguirse tras la elaboración de este trabajo de pregado.

1. Como se ha comentado a lo largo de esta memoria, la cantidad de frases traducidas a español desde la LSC es muy escasa. En este trabajo de grado se ha generado un *corpus* paralelo con frases en un solo dominio, sin embargo, interesa incrementar ese *corpus* paralelo para poder seguir mejorando los sistemas de traducción.

2. Entre las posibles formas de ampliar el *corpus*, es generar del sistema actual un sistema bidireccional, anexando al *corpus* las posibles respuestas a las frases interrogativas del actual *corpus*. El actual *corpus* corresponde a las consultas más comunes de un usuario, las nuevas frases anexadas al *corpus* corresponderían a las respuestas posiblemente dadas por el administrativo a cargo del punto de información.

3. Este sistema podría extenderse a un aula de clase. Por tal motivo, se debe ampliar el *corpus* con toda aquella terminología que tenga seña y representación en glosa en LSC y haga alusión a los conceptos de cada área del conocimiento.

4. Generar nuevos *corpus*, todos con dominio en común el contexto académico, pero que los diferencie el punto especifico al interior del centro de educación superior (Biblioteca, Cafeterias, Aulas, entre otros) y posteriormente implementar técnicas de adaptación de dominios al actual sistema de traducción.

5. Hacer una selección más detallada de los pesos de los componentes del *Translation Model*, del *Language Model*, del *Word Penalty* y del *Distortion Model* en el archivo *moses.ini* con el objetivo de encontrar los pesos ideales para este sistema de traducción.

6. Añadir un módulo que permita pasar la traducción dada en texto a voz. Esto facilitara el proceso comunicativo de las personas perilocutivas y prelocutivas.

Bibliografía

[1] Schönenberg Ávila, Gerardo, *Mediolleno: Creando oportunidades*, **'¿Quiénes son las personas con discapacidad y cuáles son sus derechos básicos?'**, Enero 5 del 2015, Recuperado de http://mediolleno.com.sv/opinion/quienes-son-las-personas-con-discapacidad-y-cuales-son-sus-derechos-basicos

[2] Servicio de Información Sobre Discapacidad, *Instituto Universitario de Integración en la Comunidad. Universidad de Salamanca*, **'Discapacidad auditiva'**, Recuperado de http://sid.usal.es/colectivos/discapacidad/discapacidades-sensoriales-expresivas/deficiencias-oido.aspx

[3] Viviendo el Sonido, *GAES*, **'Edad y Deterioro del Oido'**, Recuperado de http://www.viviendoelsonido.com/perdida_auditiva/ver/11/perdida-auditiva/sintomas/edad

[4] Organización Mundial de la Salud,**'Datos y cifras de sordera y pérdida de la audición'**, Febrero de 2017, Recuperado de http://www.who.int/mediacentre/factsheets/fs300/es/

[5] Mineducación, insor, dane, *CENSO poblacional del Departamento Administrativo Nacional de Estadística-DANE y el Registro para la Localización y Caracterización de Población con Discapacidad-RLCPD del Ministerio de Salud y la Protección Social-MINSPRO*, **'Boletín Territorial'**, 2015, Recuperado de http://www.insor.gov.co/observatorio/download/boletin_municipal/Pereira.pdf

[6] Ministerio de Educación, *Periódico Altablero No. 43*, **'Educación para Todos'**, Septiembre - Diciembre 2007, Recuperado de http://www.mineducacion.gov.co/1621/article-141881.html

[7] Achraf Othman,Mohamed Jemni, **'Statistical Sign Language Machine Translation: from English written text to American Sign Language Gloss'**, *International Journal of Computer Science Issues* Vol. 8, Issue 5, No 3, September 2011

[8] Lynette Van Zijl, *Department of Computer Science, Stellenbosch University*, **'South African Sign Language Machine Translation Project'**.

[9] López Ludeña, Veronica, **'Diseño, desarrollo y evaluación de sistemas de MT para reducir las barreras de comunicación de las personas sordas'**, *Tesis de PhD de la Universidad Politécnica de Madrid* 2014.

[10] Achraf Othman, and Mohamed Jemni, **'Statistical Sign Language Machine Translation: from English written text to American Sign Language Gloss'**, *Articulo de la Universidad de Tunez. International Journal of Computer Science Issues*, Vol. 8, Issue 5, No 3, September 2011.

[11] Jakub Kanis, Luděk Müller, **'Sign Speech Translation'**, *Articulo de la Universidad de Occidente Bohemia, Facultad de Ciencias Aplicadas y Ciberneticas*, Republica Checa, pp. 488–495. 2007

[12] Chung Hsien Wu, Hung Yu Su, Yu Hsien Chiu ,Chia Hung Lin, **'Transfer-Based Statistical Translation of Taiwanese Sign Language Using PCFG'**, *Artículo de la Universidad Nacional de Cheng Kung de Taiwan*, 2012

[13] Xiaoxue Wang, Conghui Zhu, Sheng Li, Tiejun Zhao, Dequan Zhen , **'Domain Adaptation for Statistical Machine Translation'**, *Artículo de la 12◦ Conferencia Internacional sobre Computación Natural, Sistemas Fuzzy y Descubrimiento del Conocimiento*, Pag. 1653_ 1658, 2016

[14] Marta R. Costa Jussá, José A. R. Fonollosa, **'Sistema Estadístico de Reordenamiento de Palabras en Traducción Automática'**, *Articulo del Centro de investigación TALP, Universidad Politécnica de Cataluña*, Pag. 249 _ 255

[15] Guerrero Balaguera Juan David, Pérez Holguin Wilson Javier, **'FPGA-based translation system from colombian sign language to text'**, *Artículo de la Universidad Nacional de Colombia sede Medellin*, 2014

[16] Barreto M Alex G, Amores Sonia Margarita, **'El Uso del Software de Transcripción Lingüística ELAN en el análisis de interpretación de Lengua de Señas Colombiana en el Contexto Universitario'**, *Artículo de la Universidad Pedagógica Nacional*, 2011

[17] Betancur Betancur Daniel, Gómez Vélez Mateo, Peña Palacio Alejandro, **'traducción automática del Lenguaje Dactilológico de Sordos y Sordos-mudos Mediante Sistemas Adaptativos'**, *Artículo de la Escuela de Ingeniería de Antioquia (EIA)*, 2013

[18] Manuel Ignacio Rodríguez S,Rocío del Pilar Velásquez G. , **'Historia y Gramática de la Lengua de Señas'**, *Artículo de Universidad Pedagógica Nacional e Instituto Nacional de Sordos*

[19] Lidell S., Johnson, **'American Sign Language: The phonological base'** *In Sign Language Studies, Issue 64, pages 195 _ 277.*

[20] PAULINA RAMÍREZ, **'Un breve vistazo a la educación de los sordos en Colombia'** *Lengua de señas y educación de sordos en Colombia. Santafé de Bogotá. INSOR. 1998*

[21] CENTRO VIRTUAL CERVANTES, *Diccionario de términos clave de ELE*, **'Enfoque Oral'**,1997-2017 Recuperado de https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/enfoqueoral.htm

[22] : NANCY ROZO MELO, *Portal de Lenguas de Colombia: Diversidad y Contacto*, **'La lengua de Señas Colombiana'**, Recuperado de http://lenguasdecolombia.caroycuervo.gov.co/contenido/Lenguas-de-senas-colombiana/introduccion

[23] INSTITUTO NACIONAL PARA SORDOS, **'Diccionario Básico de la Lengua de Señas Colombiana'**, 2006

[24] JOHN HUTCHINS, **'Two precursors of machine translation: Artsrouni and Trojanskij'**, Recuperado de http://ourworld.compuserve.com/homepages/WJHutchins

[25] PHILIPP KOEHN, *Universidad de Cambridge***' Statistical Machine Translation'**, 2009

[26] JOHN HUTCHINS, **'The History of Machine Translation in a Nutshell'**, Noviembre 2005, Recuperado de http://ourworld.compuserve.com/homepages/WJHutchins

[27] DANIEL JURAFSKY, JAMES H. MARTIN, **'Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition'**, 1999,

[28] C. E. SHANNON, **'A Mathematical Theory of Communication'**, 1949.

[29] P. BROWN, V. DELLA PIETRA, S. DELLA PIETRA, R. MERCER., **'The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics'**, 1999.

[30] JULIAN DAVID ECHEVERRY CORREA, **'Contributions to Speech Analytics based on Speech Recognition and Topic Identification'**, *Tesis de PhD de la Universidad Politécnica de Madrid*, 2015.

[31] PHILIPP KOEHN, *MOSES _ Statistical Machine Translation*, **'Statistical Machine Translation System: User Manual and Code Guide'**

[32] **'Competencias Profesionales Asociadas a la Traducción Automática Estadística'**, Junio 2016, Recuperado de http://repositori.uji.es/xmlui/bitstream/handle/10234/161763/TFG_2016_DolzSanchezJoan.pdf?sequence=1

[33] FRANZ JOSEF OCH AND HERMANN NEY, *Computational Linguistics*, **'A Systematic Comparison of Various Statistical Alignment Models'**, 2003, Recuperado de http://www.statmt.org/moses/giza/GIZA++.html

[34] Aingeru Mayor, Inaki Alegria, Arantza Diaz de Ilarraza,Gorka Labaka, Mikel Lersundi, Kepa Sarasola, *Procesamiento del Lenguaje Natural*, **'Evaluation of a Rule-Based Machine Tranlation system or why BLEU is only useful for what it is meant to be used'**, núm. 43 (2009), pp. 197-205

[35] Jesús Tomás Gironés, *Universidad Politécnica de Valencia*, **'Traducción Automática de Textos entre Lenguas Similares Utilizando Métodos Estadísticos'**, Noviembre del 2013

[36] Maike Erdmann, Andrew Finch, Kotaro Nakayama, Eiichiro Sumita, Takahiro Hara, Shojiro Nishio, *Workshops of International Conference on Advanced Information Networking and Applications*, **'Calculating Wikipedia Article Similarity Using Machine Translation Evaluation Metrics'**, 2011

[37] Patiño Giraldo, Luz Elena, **'La Lengua de Señas Colombiana como mediadora en el proceso de conceptualización de nociones relacionadas con las ciencias sociales en niños y niñas no oyente'**, 2010

[38] Alejandro Oviedo, *INSOR - Universidad del Valle*, **'Apuntes para una Gramática de la Lengua de Señas Colombiana'**, 2001

ANEXOS

Anexo 1: Lineas de Código para la Generación del Modelo de Lenguaje

Haciendo uso de la terminal de Linux, siguiendo los pasos establecidos en Moses y utilizando la siguiente linea:

$ROOT0/mosesdecoder/bin/lmplz -o 3 < /$ROOT1/CorpusEspanol.txt >ModeloLenguaje

Cabe resaltar que, la base de datos utilizada corresponde a la lengua destino, en este caso el *corpus* en español. Se implemento un orden de 3. $ROOT0 corresponde al *path* de la carpeta *mosesdecoder*, $ROOT1 corresponde al *path* de la carpeta donde se encuentra el *corpus* a emplear. El archivo de salida se etiqueto con el nombre de 'ModeloLenguaje' y se guardo en la carpeta donde actualmente se ejecuta la linea desde la terminal. Se obtuvo un archivo plano donde se evidencia el conteo total de unigramas, bigramas y trigramas de todo el *corpus*. Además, la escritura de cada unigrama, bigrama y trigrama con sus correspondientes pesos; los pesos se asignas según la cantidad de veces que aparezca los unigramas, bigramas y trigramas.

Anexo 2: Lineas de Código para la Generación del Modelo de Traducción y el Archivo moses.ini

La función utilizada para el alineamiento y extracción de las tablas de traducción fue **-train-model.perl** propia del *toolkit* GIZA++. Esta función tiene los siguientes parámetros:

- –root-dir : Crea un directorio en la dirección establecida por el usuario, en este directorio se guardan los archivos de salida.
- –f : Se escribe la extensión del archivo del *corpus* origen, para este caso son las glosas (.gl).
- –e : Se escribe la extensión del archivo del *corpus* destino, para este caso es el español (.es).
- –corpus: Es la dirección del *corpus* paralelo.
- –external-bin-dir : Es la dirección de ubicación de la herramienta descargada GIZA++.
- –alignment : Especifica la forma en que se realiza el alineamiento.

Para efectos de esta tesis, el parámetro para *–alignment* que se eligió fue *grow-diag-final*, que significa que la alineación comienza con la intersección de las dos alineaciones y luego agrega puntos de alineación adicionales.

Finalmente, en esta sección se generan las tablas de traducción de frase como se logra ver en la figura 7.1.

Figura 7.1: Icono que representa el archivo que contiene las tablas de traducción de frase

Lineas de Código para la Generación del Modelo de Reordenamiento

Para la construcción del modelo de reordenamiento, se utilizó la función **-reordering**, cuyos parámetros son:

- *Modeltype*: Corresponde al tipo de modelo utilizado, este parámetro puede ser definido como:
  * *wbe:*Hacer referencia a un modelo de lenguaje basado en frases y decodificador.
  * *phrase:*Corresponde a un modelo de lenguaje basado en frases.
  * *hier:*Corresponde a un modelo de lenguaje basado en frases jerárquicas.
- *Orientation*:Se refiere a que clase de orientación es usada en el modelo de lenguaje, este parámetro puede ser definido como:
  * *mslr:*Considera 4 diferentes orientaciones: *Monotone* (Monotono), *Swap* (Intercambio), *Discontinuous Left* (Discontinuo a la izquierda),*Discontinuous Right* (Discontinuo a la derecha)
  * *msd:* Considera 3 diferentes orientaciones: *Monotone*, *Swap* y *Discontinuous*[1]
  * *Monotonicity:* Considera 2 diferentes orientaciones: *Monotone* y *Non- Monotone*[2]
  * *leftright:*Considera 2 diferentes orientaciones: Derecha o Izquierda.[3]
- *Directionality*:Determina si la orientación debe ser modelada en base a la frase anterior, siguiente o ambas. Este parámetro puede ser definido como:
  * *Backward:*Determina la orientación con respecto a la frase anterior.
  * *Forward:*Determina la orientación con respecto a la frase siguiente.
  * *Bidirectional:*Determina la orientación con *Backward* y*Forward*.
- *Language*:Decide en que idioma basar el modelo, este parámetro puede ser definido como:
  * *fe:* Condiciona en ambas lenguas (destino y origen)
  * *f:*Condiciona solo en la lengua origen.
- *Collapsing*:Determina como tratar los puntajes, este parámetro puede ser definido como:
  * *Allff:*Trata los puntajes como funciones individuales.
  * *Collapseff:*Colapsar todos los puntajes en una dirección dentro de una función característica.

Para el caso particular de este trabajo de grado, los parámetros correspondientes son: para *Language* fue *fe*, para *Directionality* fue *birectional* y para *Orientation* fue *msd*. Para el caso del *Modeltype* y *Collapsing* que no se especificaron, se asumen las opciones por defecto de Moses, que son wbe y allff.

[1]Para este caso, se fusiona la discontinuidad a la derecha y a la izquierda

[2]La orientación*Swap* y *Discontinuous* son integradas en la orientación *Non- Monotone*

[3]la orientación *Swap* y *Discontinuous Left* son integradas en *Left* y *Monotone* y *Discontinuous Right* son integradas en*Right*

Generación del moses.ini

El archivo *moses.ini* es donde se especifican los pesos generados para cada modelo. En la siguiente linea se muestra los parámetros específicos utilizados para este trabajo:

```
\$ROOT0/mosesdecoder/scripts/training/train-model.perl -root-dir train
\$ROOT1/CorpusTraining -f gl -e es -alignment grow-diag-final-and -reordering
msd-bidirectional-fe -lm 0:3 \$ROOT2/ModeloLenguaje  -external-bin-dir
\$ROOT0/mosesdecoder/tools
```