Universidad Autónoma Metropolitana
Unidad Azcapotzalco
Maestría en Ciencias de la Computación

Traductor híbrido wixarika - español con escasos recursos bilingües

IDÓNEA COMUNICACIÓN DE RESULTADOS
Que para obtener el grado de
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

Presenta
Jesús Manuel Mager Hois
Matrícula: 2153801455

Asesor
Dr. Carlos Barrón Romero
UAM - Azcapotzalco

Co-asesor
Dr. Ivan Vladimir Meza Ruiz
UNAM - IIMAS

Ciudad de México — Febrero 2017

RESUMEN

Se presenta un traductor automático entre las lenguas español y wixárika, también conocida como huichol, por medio del Procesamiento de Lenguaje Natural. La lengua wixárika[1] es importante como lengua indígena ya que es hablada en los estados de Jalisco, Nayarit, Zacatecas y Durango, y tiene entre treinta y cincuenta mil hablantes. Se usa el modelo de Traducción Estadística por Frases. En el estado del arte se utilizan de entre 100MB y 300 MB de texto alineado, sin embargo, para el par de idiomas wixárika-español los textos alineados son muy escasos. Para resolver el problema se creó un analizador y segmentador morfológico que permite la separación de las palabras aglutinadas wixaritari en morfemas, lo cual permite trabajar con la polisíntesis del idioma. También se escribieron herramientas básicas para el procesamiento de lenguaje natural, como es un normalizador, para lo que se estableció un alfabeto base del idioma y un tokenizador. Estas herramientas se incorporaron a la metodología de traducción por frases. Con el fin de divulgar el trabajo realizado y de obtener retroalimentación por parte de los hablantes de la lengua, se ha creado una plataforma web donde se pueden hacer las traducciones en las dos vías, tanto español a wixárika y wixárika a español. Los resultados obtenidos son buenos al compararlos con otros trabajos de traducción, tomando en cuenta la distancia entre lenguas traducidas y los escasos recursos con los que se cuenta.

Palabras clave: Procesamiento de Lenguaje Natural, Traducción Máquina, Recursos Escasos.

[1]También conocido como huichol. Se pronuncia como wirrarica.

A mi madre, a mi familia de zoquipan y a la nación wixárika.

Agradecimientos

Agradezco a todo el pueblo mexicano que con su trabajo crea y sostiene la educación pública y gratuita de nuestro país y a todas las personas que luchan a diario por ella; a la Universidad Autónoma Metropolitana (UAM), que gracias a sus alumnos, académicos y administrativos formamos una comunidad que en conjunto alcanza la excelencia académica. También reconozco el importante esfuerzo de todos los miembros de la Maestría en Ciencias de la Computación (MCC) de la UAM, con el cual he logrado finalizar esta etapa y a la oficina de la división de CBI.

Quiero hacer mención especial a mis asesores el Dr. Carlos Barrón Romero de la UAM y al Dr. Iván Vladimir Mesa Ruiz de la UNAM; a mis revisores de tesis: Dr. Alejandro Aguilar Zavoznik (UAM), Dr. Héctor Javier Vázquez (UAM-AZC) y Dr. Jorge García Flores (LIPN-*Université de Paris 13*) y a mis profesores que me aconsejaron en el presente trabajo Dr. Raúl Miranda Tello y al Dr. Juan Villegas; al coordinador de la MCC el Dr. Luis Fernando Hoyos por su dedicación y a la directora de la división de CBI Dra. María de Lourdes Delgado Núñez por su apoyo.

También mencionaré a las personas que contribuyeron al presente trabajo: a Rebeca Guerrero con su invaluable apoyo en correcciones, transcripciones y consejos, a mi madre Dra. Elisabeth Albine Mager Hois cuyos consejos y correcciones permitieron a este trabajo mejorar su calidad, a Diónico González Carrillo, que generó y corrigió gran parte del corpus wixárika, y a Armando Martínez por el diseño del logo del traductor.

En general agradezco a mi pueblo y a mi familia wixárika, que desde mi infancia me inspiraron respeto por un mundo lejano y propio.

Índice general

Índice de figuras

Índice de tablas

Lista de algoritmos

Capítulo 1

Introducción

titayali m+	¿Por qué razón,
tiwilikuta	se llama Wirikuta?
hawelí nunuchi	Aunque soy un ser humilde
'u nelanumiet+ m+	al venir aquí
nep+timaim+k+.	mi deseo es aprender.
tisa+t+k+ t+ t+	No por nada
p+katiyetewa,	se llama así,
'alí Halamala	es a consecuencia
muwa leutimieme s+a	de lo que hizo Haramara
'alí yuchichisi	que es la madre
'alí p+wamama.	de las divinidades.

Canto wixárika, por José Bautista Carrillo y Marcos Navarrete Bautista, Tsakutsé, Tateikie. (Julio 1993)

El presente trabajo expone un sistema de traducción wixárika-español. El wixárika es un idioma hablado por entre treinta mil y cincuenta mil personas (Iturrio & Gómez López 1999), con pocos textos escritos, con un análisis gramatical limitado (Iturrio & Gómez López 1999, Grimes 1964) y sin un estudio conocido en el campo del Procesamiento de Lenguaje Natural (*Natural Language Processing*, NLP). Por ello es necesario generar las herramientas de procesamiento de este idioma, que pueden aportar además a procesar más idiomas de la familia yutonahua, a la cual pertenece y, en general a lenguajes que cuentan con pocos textos escritos. Hasta la fecha, no existe un cuerpo de NLP para el wixárika, ni un traductor automático para este caso. La aplicación del NLP a las lenguas originarias representaría un avance para incorporarlas al nuevo entorno. En

la actualidad una persona hablante de estas lenguas debe usar el español u otra lengua para poder acceder a la tecnología. Los entornos computacionales no proporcionan interfaces, correctores ortográficos o de estilo, ni contenidos en lenguas indígenas. El desarrollo del NLP permitiría acercar este mundo a la vida cotidiana de los pueblos.

El tema de la traducción es un problema duro y por lo tanto un campo propicio para la investigación. El campo semántico lleva, según Tarski, a la indefinibilidad (Tarski 1936) y, por lo tanto, los lenguajes naturales no pueden ser resueltos como lenguajes formales (de la lógica de la ciencia de la computación, por ejemplo, como un lenguaje de programación). La complejidad de la traducción automatizada debe confrontar y combinar adecuadamente los siguientes factores: las barreras culturales entre dos diferentes lenguajes, la inherente ambigüedad de los lenguajes humanos y la irregularidad entre dos lenguas (Roy 2010).

El marco teórico del trabajo incluye el procesamiento de lenguaje natural, que es un área de la ciencia de la computación que junto con el área de lenguajes de programación y compiladores han madurado a lo largo de sesenta años y se incluyen técnicas de reconocimiento de patrones y *data mining*. Todo lo anterior se combina con una modelación matemática adecuada del problema particular.

Al diseño, análisis y recopilación de datos para el traductor entre los dos idiomas se aplicarán los algoritmos y métodos usados en otras lenguas más investigadas, cómo, náhuatl y turco (Ermolaeva 2014). El problema de la traducción ha sido investigado para lenguas que generalmente cuentan con un gran cuerpo de documentos apareados para su traducción. Sin embargo, en este caso – wixárika y español – se tiene un conocimiento gramatical adecuado, pero pocos documentos apareados, por lo que es necesario adecuar las metodologías de frontera de NLP y modelar adecuadamente.

La posibilidad de que un idioma como el wixárika cuente con un traductor automático, permitirá la traducción de libros y otro tipo de contenidos escritos. Esto facilitará a sus hablantes poder acceder a los textos en su propio idioma. Sin embargo, el presente trabajo se centra en plantear traducciones simples, como un primer paso para el problema de traducción de lenguas indígenas.

1.1. Justificación

En México se hablan sesenta y ocho lenguas originarias (de Lenguas Indígenas 2016) de las cuales veintiún cuentan con menos de mil hablantes. La UNESCO identificó en el año

2007 que el cincuenta por ciento de las lenguas a nivel mundial se encuentran en peligro de desaparecer, seis mil lenguas son habladas únicamente por el cuatro por ciento de la población mundial y el noventa por ciento de las lenguas no están representadas en Internet (UNESCO 2007). Esto plantea un problema muy importante y trascendente para la cultura universal y los valores humanos, que no es exclusivo de nuestro país: la preservación de la cultura y las lenguas indígenas.

La comunicación es fundamental para la interacción en la sociedad y el lenguaje es la principal vía por la cual se realiza esta comunicación y funge además como un factor para la unidad de los grupos étnicos, representando el papel de "*lenguaje común*, que es el pensamiento mismo y constituye un código compartido, un campo semántico elaborado históricamente, según el cual se organiza la comprensión del mundo" (Guillermo 1981). Este cuerpo semántico es propio y es el que hace que cada lengua guarde historia y conocimientos. Por lo tanto, cada vez que desaparece un lenguaje, la humanidad pierde una parte de su semántica universal y parte del *patrimonio cultural inmaterial*[1].

Son los pueblos originarios los que más han sufrido el impacto del uso de las Tecnologías de la Información y la Comunicación (TIC) ya que la transmisión de las lenguas indígenas es oral, de generación en generación. Algunas no cuentan con lenguaje escrito y no utilizan la tecnología o si lo hacen son afectados por la obligación de utilizar lenguas dominantes, como por ejemplo el español o el inglés. Además existe una falta de profesores bilingües en el sistema educativo mexicano y los docentes con estas características están son formados para llevar acabo una castellanización forzada de las comunidades, según Martínez (Martínez Casas 2011). La falta de una eseñanza de la lengua escrita, y su supresión en el salón de clase afecta su superviviencia y futuro, lo que aunado a tecnologías en español como el Internet, se convierte en la aniquilación de los lenguajes que "no son útiles". Esto impulsa cada vez más la extinción de las lenguas poco habladas como en el pasado fue la tendencia de que los primeros intentos de comunicación entre culturas distintas consistía en imponer la religión y cultura dominante, a través de traducciones de la biblia (ver los trabajos del Lingüístico Verano (Grimes 1964)).

El trabajo propone una interacción más libre y justa mediante una herramienta automática de traducción de uso general, en las dos vías (wixárika a español y español a wixárika). Eso permite a los hablantes, de las dos lenguas, elegir los textos y los temas de su interés, intercambiando conocimientos de acuerdo a sus propias necesidades e

[1]Una de las manifestaciones particulares del "patrimonio cultural inmaterial" son tradiciones y expresiones orales, incluido el idioma como vehículo del patrimonio" (UNESCO 2003).

intereses.

1.2. Objetivos

Partiendo del marco del NLP y en particular de la traducción automatizada aplicada al caso de la traducción con el par de idiomas wixárika-español. El desarrollo se realizará con un diseño en espiral para mantener adaptabilidad. El sistema podrá mejorar ya que los datos dirigen parte importante de la traducción. El problema al ser duro, y no tener solución definitiva incluso para los pares de idiomas más estudiados pretende continuar aprendiendo a partir de la información proporcionada por los usuarios. El trabajo es un caso de estudio de traducción automatizada para un par de lenguajes que no cuentan con una gran cantidad de frases emparejadas: traducción del español al wixárika y viceversa. Para ello se adaptarán las herramientas del NLP al traductor. Si bien será un estudio de un caso particular y pretende expandirse a reglas de otros idiomas, las conclusiones permitirán apoyar a idiomas con dificultades semejantes y convertirse en aportes universales. Por lo tanto, la hipótesis del trabajo es que un sistema puede generar traducciones automáticas entre el wixárika y el español.

Objetivo: Construcción de un sistema de traducción híbrido wixárika-español en dos vías, que sea la base del NLP en wixárika.

Metas:

- Construir una base de datos de frases emparejadas para traducción automática entre wixárika y español. (ver apéndice C)
- Generar una herramienta identificadora del wixárika (ver apéndice A).
- Crear un tokenizador, normalizador y un analizador morfológico, que ayudarán a la herramienta traductora (ver apéndice A).
- Implementar un diccionario (ver apéndice B).
- Crear una plataforma web en la cual se expondrá el diccionario, se podrá aportar frases wixáritari-español y el traductor mismo (ver sección 3.6).
- Servirá para la asistencia de traducción humana.

1.3. Metodología

Para poder cumplir con los objetivos y metas se utiliza el método inductivo-deductivo con las siguientes etapas:

1. **Investigación bibliográfica y hemerográfica. (Marco teórico)**

   Realizar una investigación de los diferentes algoritmos de traducción, los paradigmas existentes y las discusiones entre ellos para experimentar con los algoritmos más prometedores.

2. **Recolección de materiales.**

   Tomando el planteamiento de que los algoritmos de traducción estadística han generado los mejores resultados, se debe buscar una máxima cantidad de frases o textos wixárika apareados con el español. Entre mayor sea la cantidad de datos encontrados mayor será la probabilidad de generar buenas traducciones.

3. **Modelación y diseño.**

   Se creará un diseño del modelo de traducción a utilizar y su interacción con los mecanismos auxiliares como redes semánticas, generación de diccionarios con vectores de relación, además del diseño de una interfaz hombre-máquina y el diseño de métodos de evaluación de los resultados.

4. **Programación y desarrollo del sistema.**

   Antes de poder utilizar el corpus emparejado se necesita preparar los textos recolectados y hacerlos aptos para su procesamiento posterior. Una vez encontrada la información, se podrán usar las frases de entrenamiento bilingües recabados en la fase de recolección de materiales y usar sistemas existentes como Moses (Moses 2016).

5. **Análisis de los resultados.**

   Con base en los resultados obtenidos con las diversas estrategias de traducción, se generarán razonamientos inductivos para mejorar el comportamiento del caso particular de los idiomas que se pretenden traducir.

6. **Adaptación y evaluación.**

Dependiendo de los resultados obtenidos se realizará una adaptación de los modelos con mayor expectativa de generar buenas traducciones para el wixárika-español. Se repetirán los pasos 4, 5 y 6, para generar resultados satisfactorios.

7. **Redacción de los resultados.** Se presentará el trabajo en congresos especializados y se realizará la Escritura de la Idónea Comunicación de Resultados. Presentación del examen de grados.

El desarrollo de un traductor automático al wixárika será la primera aplicación de NLP al wixárika, por lo que existe un amplio terreno para investigaciones futuras. En el estudio de lenguajes originarios se han realizado algunos trabajos como el corpus Axolotl de Ximena Gutiérrez del náhuatl (Geographic 2016) y el proyecto *Microsoft Translator Community Partners* (Microsoft 2016), para el ñañú (otomí) Querétaro y el maya de Yucatán. Este último proyecto de Microsoft no es libre o abierto, por lo que no se puede acceder a sus avances.

El presente trabajo se organiza en tres capítulos de la siguiente forma:

- **Estado del arte**

  El capítulo presenta un recuento de la historia de la traducción automática, expone los principales paradigmas de traducción de la actualidad y por último describe los fundamentos de la traducción automatizada estadística: la traducción basada en palabras y la basada en frases.

- **Metodología**

  Se expone la metodología utilizada para tratar el problema de la traducción con bajos recursos y de una lengua aglutinante. Se presenta el idioma wixárika y el segmentador morfológico.

- **Resultados**

  Con base en lo discutido en los capítulos previos, se presentan los resultados de una prueba de concepto, y la traducción en dos sentidos, tanto de wixárika a español, como de español a wixárika. Por último se comparan los resultados obtenidos con el modelo propuesto con casos de traducciones semejantes.

El trabajo se limita a un traductor estadístico básico que tenga capacidad de aprender con un sistema de incorporación de corpus nuevo a través de una interfaz web.

Podrá generar una traducción inicial de frases sencillas, sobre temas específicos. El sistema tendrá la capacidad de ser extendido y mejorado en estudios futuros. En el triángulo de Vauquois se pretende llegar al nivel sintáctico, pero no se descarta el posible uso de técnicas del nivel semántico.

Capítulo 2

Estado del Arte

El campo de conocimiento de la ciencia de la computación conocido como procesamiento de lenguaje natural permite el estudio de diversas tareas entre los lenguajes naturales y las computadoras, como son el reconocimiento de voz, la generación de textos, la extracción de información, la traducción automática, entre otras. Sin embargo, el NLP se ha centrado en los lenguajes más hablados y existen pocos ejemplos para lenguajes con pocos textos escritos. En el caso de la traducción automática casi no existen traductores automáticos para lenguas originarias, pero si han sido extensamente trabajadas para alemán, español, francés, italiano, portugués, árabe, japonés, koreano, chino, holandés, griego y ruso (en sistemas comerciales y públicos como Google, Systran, Prompt); y en casi todos los casos el inglés es la contra parte de las traducciones (Laukaitis & Vasilecas 2007).

2.1. La traducción y la traducción automática

Para poder comenzar es necesario definir conceptos básicos del tema a tratar, entender que es un lenguaje y una traducción. Se explicará posteriormente la historia de la traducción automática y finalmente se presentará una introducción al idioma wixárika en sus propiedades sintácticas y morfológicas.

2.1.1. Definiciones

Un alfabeto es un conjunto no vacío de símbolos, y se utiliza el símbolo Σ para representarlo. El idioma inglés tiene 26 símbolos en Σ, el español 33, sin considerar símbolos

de puntuación, espacios y mayúsculas. El alfabeto es adquirido por cuestiones históricas, tanto por desarrollo propio, por adopción, o por imposición. El Σ del wixárika fue desarrollado por los misioneros españoles que trabajaron en la zona y posteriormente por lingüistas y la SEP. En la tabla 3.1 se muestra el alfabeto utilizado para el wixárika.

Una palabra es una secuencia finita de símbolos pertenecientes a Σ. En este texto se denota una palabra como w. Las palabras tienen un tamaño según el número de símbolos que contengan, denotado como $|w|$. Una palabra de tamaño cero es denotado como ϵ. La palabra vacía es usada para presentar una correspondencia nula de otra palabra, o a la inexistencia de una palabra en el idioma relacionado, entre otros. A pesar de contar con una gran claridad de qué es una palabra formalmente, en el contexto de los lenguajes naturales, esto no es tan claro. Existen idiomas donde las palabras se distinguen fácilmente, como es el caso del español o del inglés, sin embargo, en idiomas como el chino, su escritura no delimita entre una palabra y otra.

Σ^+ es un conjunto de todas las palabras que puede generar un alfabeto Σ, excepto la palabra vacía. El conjunto de todas las palabras, incluyendo la palabra vacía, sobre un alfabeto es conocido como cerradura de Kleene y se denota por Σ^*. Así se obtiene $\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$ y $\Sigma^* = \Sigma^+ \cup \{\epsilon\}$(Hopcroft et al. 2000).

Ahora bien, se define a un lenguaje L cómo un conjunto de palabras sobre un alfabeto Σ, tal que $L \subset \Sigma^*$. El lenguaje natural, en contraste con los lenguajes formales, tiene dos campos, el semántico y el sintáctico. El campo semántico es el de mayo complejidad, y es en este campo donde se logra encontrar el verdadero significado de una frase o un texto.

Una traducción es una actividad que a partir de un texto en un lenguaje origen genera un texto en un idioma meta. El significado del lenguaje meta tiende a ser equivalente al del idioma origen, pero no lo es completamente, al perderse y enriquecerse con contextos semánticos y particularidades gramaticales de cada lenguaje. Al texto traducido, se le denomina como traducción y se denota como e, mientras que el texto origen es denotado como f. La ciencia que estudia la traducción es llamada traductología, y se centra en la teoría, descripción y la aplicación de la traducción.

2.1.2. Historia

Para poder entender el problema de traducción de una manera más amplia, se presenta un esbozo histórico de su desarrollo. En la Tabla 2.1 se muestra una linea del tiempo de la traducción automatizada. La traducción, tanto verbal como escrita, ha servido desde

el surgimiento de la humanidad para facilitar la comunicación entre los pueblos, pero también para imponer o preservar ideas, conquistar nuevas tierras u obtener secretos. La humanidad ha tenido una larga historia de traducción, donde el primer registro data del año 196 a.n.e., con la escritura de una estela egipcia, conocida como *piedra rosseta*. Se encontraron tres idiomas, donde el texto superior estaba escrito con jeroglíficos egipcios, el segundo texto en escritura demótica y, por último, la inferior, en griego antiguo. Estos textos permitieron avanzar en descifrar la escritura jeroglífica egipcia. Si bien este primer registro ayudó a cuestiones administrativas del imperio egipcio, las traducciones también se han utilizado para expandir religiones y las visiones del mundo.

Si bien, la traducción humana entre dos lenguajes ha sido utilizada por milenios, es imposible traducir con exactitud un texto a otro lenguaje. Esto se comprueba en el teorema de la Indefinibilidad de la Verdad de Tarski, donde se plantea que un lenguaje natural no formalizado es más expresivo que uno restringido, lo que permite que uno solo de sus elementos tenga múltiples significados. Al momento de tomar el concepto de verdad, se podría derivar del mismo su propia negación, y nos permite entender el motivo por el cual no es posible traducir sin pérdida de información (Barron et al. 2016).

A pesar de que la traducción no puede ser expresada como una relación uno a uno, a partir de los avances en criptología y codificación se comenzaron a hacer grandes esfuerzos por adentrarse en la traducción automatizada. Durante la segunda guerra mundial, los científicos búlgaros e ingleses lograron descifrar el código alemán *Enigma* con la ayuda de la computación. A partir de este gran éxito la traducción entre dos lenguajes se percibió como un proceso de decodificación de una lengua extranjera. Este punto de vista, evidentemente, no contemplaba la complejidad inherente de la semántica humana, pero fue un primer incentivo para comenzar con los esfuerzos en el campo.

En el inicio de la traducción automática se comenzaron a probar diferentes metodologías, que fueron desde una simple traducción directa, palabra a palabra, usando algunas reglas simples, hasta métodos más refinados que utilizaban análisis semántico y morfológico. El mayor interés fue mostrado por las instituciones de seguridad estatales de Estados Unidos, que centraron toda su atención en la traducción del ruso al inglés, dada la confrontación mundial entre comunismo y capitalismo, en el mundo posterior a la Segunda Guerra Imperialista Mundial. El 7 de enero de 1954, se presentó el exp:erimento de Georgetown-IBM, desarrollado por la Universidad de Georgetown e IBM, con un vocabulario de 250 palabras y 6 reglas gramaticales. Se lograron traducir con éxito, y de manera completamente automática, sesenta oraciones del ruso al

inglés(Hutchins 2004). Esto atrajo un mayor financiamiento por parte del gobierno estadounidense, e incrementó los esfuerzos en las siguientes décadas. El objetivo que se planteó la comunidad fue una traducción totalmente automática de alta calidad.

Año	**Suceso**	**Autor**	**Comentario**
196 a.n.e.	Piedra Rosetta	Ptolomeo V	Primer registro de traducción en tres idiomas
Siglo II d.n.e.	Traducción de la Biblia del griego al latín	.	Es el inicio de una fuerte difusión de la biblia en el mundo
Siglos IX y X	Textos clásicos griegos al árabe	Bagdad	Hubo expansión de los avances científicos y la cultura helénica y occidental
1939	Descifrado de Enigma	Inteligencia Polaca e Inglesa	
1954	Experimento	Universidad de Georgetown e IBM	Primera demostración de traducción automática
1966	Reporte ALPAC	ALPAC	Demuestra que no se lograron alcanzar los avances esperados en TA
1968	Systran	Peter Toma	Primer traductor basado en reglas comercializado (Inglés – Ruso)
1976	Météo	TAUM	Sistema de traducción de informes meteorológicos
1988	Traducción Estadística CANDIDE	Grupo de investigación IBM	Se presenta la traducción estadística en un modelo por palabras.
1993	Verbomil	Ministerio Federal de Investigación alemán	Sistema de traducción simultánea basada en Ínterlengua
1999	Traducción por frases	Och, Tillman, Ney	Se plantea el primer modelo de traducción estadística por frases

Tabla 2.1: Desarrollo histórico de la traducción automática

Sin embargo, en 1966, una década después de grandes esfuerzos y recursos invertidos en el tema, se presentó el reporte ALPAC (*Automatic Language Processing Advisory Committee*). El reporte demostró que los resultados obtenidos después de una revisión de un texto traducido por una máquina no eran más baratos que una traducción humana, que se contaba con suficientes traductores humanos y que existían muy pocos textos que se deseaba traducir(Koehn 2010). Con ello, era claro que la meta de conseguir una traducción de alta calidad de manera automática aún estaba distante, o incluso imposible. Con estos resultados se perdió interés en el tema, lo cual no impidió que se siguiera trabajando en él.

Con el desalentador horizonte planteado, se continuó la experimentación y la investigación para crear un sistema completo de traducción. Los primeros sistemas comerciales fueron presentados en los años setenta, a pesar de las desalentadoras perspectivas que presentó el informe ALPAC. El primero en su tipo fue *Systran*, fundado en 1968 por Peter Toma.Fue usado desde 1970 por la Fuerza Área de Estados Unidos y en un inicio únicamente traducía del ruso-inglés (Koehn 2010). La Comisión Europea también adquirió una versión, esta vez inglés-francés, con lo que se comenzaron a desarrollar más pares de idiomas. Systran es un traductor basado en reglas y en la actualidad cuenta con cuarenta pares de idiomas, es multiplataforma y sigue desarrollándose. En 1976, se presentó el sistema MÉTÉO por el grupo TAUM (*Traduction Automatique de l'Université de Montréal*), desarrollado para traducir informes meteorológicos en Inglés-Francés y fue usado desde 1982 hasta el 2001.

En la décadas de los ochenta y noventas la investigación se centró en el desarrollo de sistemas ínterlingua, que llevaría a una formalización de los significados, por lo que se trabajó en una teoría formal del conocimiento, que es uno de los grandes retos del aprendizaje máquina y de la filosofía. Con la extensión de la formalización de la gramática, se trabajó en la creación de una manera de expresar las dos partes del conocimiento: la parte *paradigmática* y la *sintagmática*. Estas serán representadas mediante un lenguaje intermedio, que permite relacionar los significados (Bahattacharyya 2015). Como ejemplos importantes de traducción usando interlingua tenemos: el traductor CATALTYST de la Universidad Carnage Mellon, y el proyecto *Verbmobil* (Wahlster 1997) desarrollado entre 1993 y 2000, donde se intentó, por parte del Ministerio Federal de Investigación alemán, crear un sistema que pudiese traducir una conversación espontánea de manera robusta y bidireccional para el Alemán-Inglés y el Alemán-Japonés.

En los años ochenta surge el concepto de métodos de traducción impulsados por datos, con los primeros intentos realizados por traducción basada en ejemplos. En los

laboratorios de IBM, surgió el modelo de una traducción estadística (Brown et al. 1988), inspirándose en los métodos estadísticos de reconocimiento de voz que estaban dando sus primeros pasos. Sin embargo, en ese momento no tuvo mayores repercusiones, al estar el paradigma centrado en los sistemas basados en reglas e interlingua. El sistema que se desarrolló fue CANDIDE (Berger et al. 1994), que fue el primer sistema estadístico basado en palabras. En 1998, los participantes en un taller de la Universidad de Johns Hopkins implementaron la mayoría de los modelos IBM (Brown et al. 1988) e hicieron públicas sus herramientas, lo cual permitiría la experimentación de más personas en el modelo, llevando a un rompimiento del paradigma imperante. Con los trabajos de Och, Tillman y Ney (Och et al. 1999) se comenzó la etapa de la traducción basada en frases. El sistema más emblemático en software libre es Moses(Moses 2016); pero también los traductores comerciales Bing y Google Translate funcionan con este paradigma.

Desde entonces se ha trabajado en los dos principales paradigmas, la traducción estadística y la traducción basada en reglas, además de modelos híbridos entre los anteriores. Con los atentados terroristas del 11 de Septiembre de 2001 en Estados Unidos y los recientes conflictos bélicos, se ha revivido el interés en financiar proyectos de traducción automática, sobre todo en el par de lenguas inglés - árabe y ruso - inglés(Koehn 2010).

2.2. Modelos de traducción automática

Como se ha visto en la sección anterior, el desarrollo de la traducción automática (*Machine Translation*) se movió por momentos y tendencias. Se mostrará ahora los modelos más relevantes en la actualidad, los que se pueden dividir estas en tres campos: la traducción basada en reglas (RBMT), los modelos estadísticos (SMT) y la traducción basada en ejemplos (EBMT) (Bahattacharyya 2015). A continuación se explican estos modelos.

- **RBMT** En el triángulo de Vauquois (Jurafsky & Martin 2000)(que se muestra en la figura 2.1) se explica cómo se logran relacionar diferentes niveles de reglas de traducción entre dos lenguajes. Existen reglas que definen el análisis de los enunciados origen, reglas de cómo transferir las representaciones y finalmente reglas para generar texto de la representación transferida (Bahattacharyya 2015). Este proceso es conocido como análisis–transferencia–generación (ATG). En el caso de que sus reglas sean aplicadas exactamente al caso de traducción, el resultado

será de alta calidad y muy preciso, con la ventaja de poder explicar el resultado de la traducción. Pero no es frecuente que sus reglas apliquen a los casos analizados, con conflictos de reglas o múltiples reglas aplicadas en un mismo caso (Bahattacharyya 2015). Dado que no existe transferencia exacta entre un lenguaje y otro,

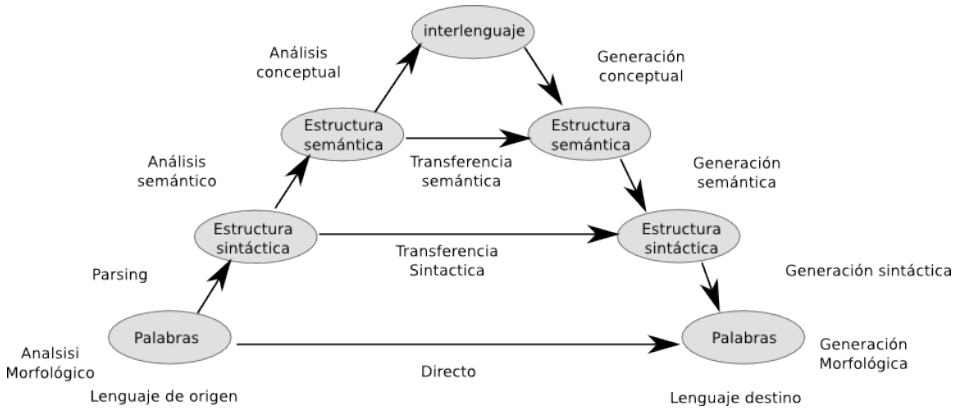

Figura 2.1: Triángulo de Vauquois

y que la sintaxis de un enunciado también depende de un contexto semántico, no es posible aspirar a poder completar el sistema de reglas que logre abarcar todos los casos con una traducción perfecta entre dos idiomas.

- **SMT** En la traducción automática estadística (SMT) las reglas de traducción ATG no son creadas a priori usando los conocimientos lingüísticos, sino que son generados a partir de un conjunto de textos emparejados. Las reglas y palabras son aprendidas de los datos de entradas y son traducidos basados en probabilidades (Bahattacharyya 2015). Estos modelos requieren un gran número de datos para poder funcionar correctamente.

- **EBMT**

  EBMT es considerado como un modelo intermedio entre SMT y RBMT. Utiliza reglas basadas en conocimientos y datos para realizar las traducciones. Los patrones de traducción provienen de los datos, pero en gran medida se utilizan reglas para determinar estos patrones (Bahattacharyya 2015).

2.3. Traducción Automática Estadística (SMT)

En este trabajo se ha planteado tomar el camino de la SMT para el traductor wixárika-español. Es por ello que se expondrá los conceptos fundamentales de SMT. Gran parte de los aportes utilizados en el modelo basado en palabras (WBSMT) serán utilizados en el modelo por frases (PBSMT), por lo que se expone las partes importantes del WBSMT, para posteriormente explicar PBSMT.

2.3.1. Planteamiento del problema SMT

El modelo que se va a utilizar para este trabajo es la traducción automática estadística, por lo que se presenta el planteamiento general del mismo. Se toman dos frases, la primera en idioma meta e y una en un idioma origen f. Para cada par de frases (e, f) existe un $Pr(f|e)$, que es la probabilidad que un traductor con entrada e produciría f. Por el teorema de Bayes se puede reformular lo anterior de la siguiente manera:

$$Pr(e|f) = \frac{Pr(f|e)Pr(e)}{Pr(f)} \tag{2.1}$$

$$= Pr(f|e)Pr(e). \tag{2.2}$$

Dado que el denominador es independiente de e se procede a tomarlo como una probabilidad de uno. Ahora bien, como se busca estimar la mejor $\hat{e}$ se utiliza el criterio de la máxima verosimilitud que se expresa de la siguiente manera:

$$\hat{e} = \text{argmax}_e p(e|f) \tag{2.3}$$

$$= \text{argmax}_e p(f|e)p(e). \tag{2.4}$$

El término $p(f|e)$ nos indica la probabilidad de que f sea el resultado del canal ruidoso cuando e es la entrada, y se conoce como el *modelo de transición*. Su dominio son todos los pares (f, e). El término $p(e)$ modela la probabilidad a *a priori* de e y es llamado el modelo del lenguaje, y desde ahora se va a expresar como $p_L M(e)$ y es utilizado para corregir el texto de salida, asegurándose que la salida corresponda a la gramática del lenguaje objetivo. Cada uno de los dos factores produce una evaluación para la frase e, donde se busca es maximizar esa evaluación. Sin embargo, la pregunta

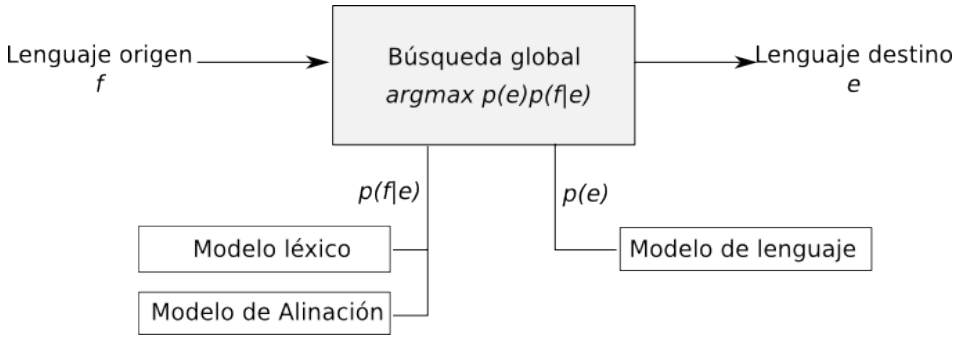

Figura 2.2: Modelo de traducción por palabras, (Zens et al. 2002)

central es como generar estos dos modelos (Brown et al. 1993).

2.3.2. Traducción basada en palabras

Como se ha mencionado, en los años ochenta del siglo pasado, el proyecto *Candide* de IBM(Brown et al. 1988) fue el primer traductor estadístico y se basó en palabras (Koehn 2010). Este modelo ya no es parte del estado del arte, pero varias de sus técnicas aún son usadas hoy en día. La traducción se basa en la probabilidad de que dada una palabra en el origen corresponda a una palabra en la meta. Con una cantidad de datos apareados, esta probabilidad será el número de veces que aparecen las palabras meta cuando aparece la palabra origen en el mismo enunciado emparejado. Este modelo se basa en la probabilidad de traducir un enunciado de un idioma fuente a un enunciado meta, con un alineamiento de cada palabra f_i con e_j, de acuerdo a una función de alineamiento $a : j \rightarrow i$,

Para poder hacer la búsqueda de la mejor traducción, descrita en 2.2 se requieren dos elementos: la estimación del modelo de alineación, léxico y un modelo de lenguaje. A continuación se profundiza en estos modelos.

2.3.3. Alineamiento

Para poder llevar a cabo la traducción de la mejor forma se considera que las frases e y f se descomponen en palabras y que las palabras de las frases tienen una correspondencia entre ellas, lo cual no es un problema menor. Si tomamos en cuenta que incluso para un humano esta es una tarea difícil, donde incluso no existen correspondencias. Sean $f^J = (f_1, \ldots, f_j, \ldots, f_J)$ una frase origen compuesta por una tupla de palabras f_j y $e^I = (e_1, \ldots, e_i, \ldots, e_I)$ una frase objetivo compuesto por palabras e_i, entonces se

"define una alineación entre dos palabras como un subconjunto del producto cartesiano de la posición de las palabras"(Och & Ney 2003).

Se denota el conjunto de alineamiento $(f|e)$ por $\mathbb{A}(e, f)$. Si e tiene un tamaño L y f un tamaño J, entonces existen LJ diferentes conexiones posibles, y se define de como $\mathbb{A} \subset \{(i, j) : j = 1, \ldots, J; i = 1, \ldots, I\}$ (Och & Ney 2003), donde $i = j \vee i \neq j$. Los alineamientos $i = a_j$ pueden contener una palabra vacía e_0. Si se supondría que una palabra f_i tiene una única palabra alineada en e_j o e_0 se obtendría una función de alineamiento $j \to i = a_j$ y no una relación.

Retomando la formula 2.4, agregando el hecho de que la composición de las frases e y f contienen palabras obtenemos $Pr(f^J|e^I)$. A esta probabilidad se introduce un factor de alineamiento oculto a_1^J que describe una función desde una posición j a una posición a_j.

$$Pr(f^J|e^I) = \sum_{a^J} Pr(f^J, a^J|e^I). \tag{2.5}$$

Ahora se va a definir a θ como un conjunto de parámetros desconocidos, que necesita un modelo estadístico y que se aprenderán de los datos en el entrenamiento $p_\theta(f^J|e^I) = Pr(f^J|e^I)$. Ahora, sean $S = \{(f_s, e_s) : s = 1, \ldots, S\}$ un conjunto de frases alineadas de un corpus paralelo, para cada par alineado se encuentra el valor de θ buscando la máxima verosimilitud

$$\hat{\theta} = \text{argmax}_\theta \prod_{s=1}^{S} \sum_{a} p_\theta(f_s, a|e_s). \tag{2.6}$$

Para cada enunciado existe una gran variedad de alineamientos $\hat{a}$, pero se tratará de encontrar el mejor alineamiento, también llamado alineamiento *Viterbi*, tal que

$$\hat{a}^J = \text{argmax}_{a^J} p_\theta(f^J, a^J|e^I). \tag{2.7}$$

2.3.4. Modelos de alineamiento

Como se ha mostrado, se requiere un modelo de alineamiento para poder llevar acabo la estimación de la máxima esperanza. Ahora se presentarán los seis modelos de alineamiento que han sido desarrollados, principalmente por Brown (Brown et al. 1993)

y el equipo de *Candide* (Berger et al. 1994). Sin entrar a los detalles matemáticos, únicamente se presenta su formulación y las ventajas o desventajas con respecto al anterior.

IBM-1

En este modelo se parte de una probabilidad de traducción léxica de una palabra del idioma origen a la meta. La traducción léxica es la probabilidad de que una palabra en el idioma origen se traduzca como otra en el idioma meta. Con estas probabilidades se obtendrán traducciones con diferentes probabilidades. Los enunciados se dividen en subproblemas de traducción, donde el problema se plantea a nivel de palabras, convirtiendo IBM-1 (Brown et al. 1988) en un modelo generativo. El modelo de IBM-I se expresa matemáticamente de la siguiente manera:

$$p(e, a|f) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)}), \tag{2.8}$$

donde el centro de la función se basa en el producto sobre las probabilidades de traducción de todas las l palabras e_j generadas en la salida, mientras que la primer parte es usada para la normalización de la función. El primer término es utilizado para normalizar, y es $l_f + 1$ por la inclusión al modelo de la palabra vacía. El parámetro ϵ es una constante de normalización, por lo general de tamaño marginal.

IBM-2

En el modelo IBM-1 únicamente se tomó en cuenta la traducción entre palabras, sin considerar su posición en los enunciados, en cambio en el modelo IBM-2 (Brown et al. 1988) se incorpora al modelo la alineación y la posición de las palabras en las frases, donde la posición j de una palabra en f^J corresponde a una posición en la frase meta i, de modo que tenemos $a(j|j, l_e, l_f)$.

$$p(e, a|f) = \epsilon \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) a(a(j)|j, l_e, l_f). \tag{2.9}$$

Por lo comentado, ahora se tienen dos etapas, donde primero se genera la traducción

léxica y posteriormente se pasará a la fase de alineamiento.

IBM-3

El problema con los modelos IBM-1 e IBM-2 es el hecho de que únicamente pueden realizar un alineamiento de una palabra a una palabra. Sin embargo, entre dos idiomas existen muchos casos donde una palabra de un idioma se expresa con dos o más palabras en el idioma meta. Los modelos anteriores asignan únicamente una palabra a la palabra origen, y las palabras sobrantes serán asignadas con la palabra vacía.

En IBM-3 e IBM-4 (Brown et al. 1988) se propone un modelo de *fertilidad*, que se expresa como $n(\phi|f)$. Este modelo nos indica que por cada palabra en f, a cuántas palabras comúnmente se traduce en e. Esto puede ejemplificarse mediante la palabra origen *neki* en wixárika, donde su fertilidad en una traducción al español sería $n(2|neki) \simeq 1$, ya que su traducción sería *mi casa*, utilizando dos palabras en español. Los posibles valores de fertilidad para una palabra pueden ser $0, 1, 2, \ldots$.

Ahora bien, para generar el modelo final, se requieren cuatro pasos, a diferencia de los dos pasos de IBM-2 y el único paso de IBM-1. En el primer paso se calcula el modelo de fertilidad, en el segundo paso se pasa a la asignación de las palabras vacías, en el tercero se genera la traducción léxica con la distribución de probabilidad $t(e|f)$ y por último, el modelo de distorsión, que es, en lo general, lo mismo que la alineación de IBM-2, pero tomando en cuenta la posición de origen y múltiples posiciones de salida $d(j|i, l_e, l_f)$.

El modelo de fertilidad plantea como $n(\phi_i|f_i)$. "Las palabras vacías ϕ_0 dependen del número de palabras de salida generadas por las palabras de entrada, y cada una de ellas puede insertar un token nulo. Por lo tanto existen $\sum_{i=1}^{l_f} \phi = l_e - \phi_0$ palabras generadas por palabras origen"(Koehn 2010). La probabilidad de generar ϕ_0 palabras de la palabra vacía es

$$p(\phi_0) = \binom{l_e - \phi_0}{\phi_0} p_1^{\phi_0} p_0^{l_e - 2\phi_0}. \tag{2.10}$$

Combinando la inserción de palabra vacía y fertilidad se obtiene

$$\binom{l_e - \phi_0}{\phi_0} p_1^{\phi_0} p_0^{l_e - 2\phi_0} \prod_{i=1}^{l_f} \phi_i! n(\phi_i | f_i). \tag{2.11}$$

La combinación de la fertilidad, la distorsión y la transferencia léxica, se permite expresar mediante la formula 2.12. En ella, el primer elemento es la fertilidad y el segundo término combina la transferencia y la distorsión. Pero, a diferencia de los modelos IBM-1 y 2, en este modelo nos encontramos con el problema de no poder reducir la complejidad del espacio de alineamientos posibles, el cual es exponencial:

$$p(e|f) = \sum_a p(e, a|f) \tag{2.12}$$

$$= \sum_{a(1)=0}^{l_f} \sum_{a(l_e)=0}^{l_f} \binom{l_e - \phi_0}{\phi_0} p_1^{\phi_0} p_0^{l_e - 2\phi_0} \prod_{i=1}^{l_f} \phi_i! n(\phi_i | f_i) \times \prod_{j=1}^{l_e} t(e_j | f_{a(j)}) a(a(j)|j, l_e, l_f). \tag{2.13}$$

IBM-4

El modelo 3 no funciona muy bien con frases largas de entrada o salida, y se obtendrán datos de salida dispersos. Los datos que están juntos en la entrada suelen estar cerca en la salida. En IBM-4 (Brown et al. 1988) se agrega un modelo de distorsión relativa. La posición de destino estará relacionada con la palabra anterior. Esto se complica si tomamos en cuenta que las palabras podrán ser desechadas o tienen una relación de una a varias.

Cada palabra f_i que es alineada a por lo menos una palabra salida es un septo. El conjunto de septos es denotado por π_i. Además se define el operador $[i]$ para mapear el septo con índice i a la posición correspondiente en f. El centro de un septo es definido como la media de posiciones de salida para un septo. Se denota por $\odot$.

Para la esperanza de asignaciones de la primer palabra de un septo se utiliza la distribución $d_1(d - \odot_{i-1})$. La asignación es la posición i de la frase destino relativo a $\odot$. Para las palabras subsecuentes del septo se utiliza la distribución $d_{>1}(j - \pi_{i,k-1})$. El valor de $\pi_{i,k-1}$ se refiere a la $k$esima palabra en el $i$esimo septo. De esta manera, es posible mejorar el modelo IBM-4.

IBM-5

A pesar de los aspectos que cubren los modelos IBM-3 y 4, se detectó un problema en ellos. Acomodar en la misma posición de salida varias palabras es imposible en la realidad, y es denotado como deficiencia, por parte del grupo de IBM. Los dos modelos mencionados no evitan la deficiencia, lo cual representa un problema. Lo anterior resulta en que alineaciones imposibles obtengan probabilidades positivas. Un efecto secundario, identificado por Och (Och & Ney 2003), es una tendencia a asignar palabras a la palabra vacía, y esto resulta en un una mala calidad de traducción. Se identifica que las palabras no vacías tienen una fertilidad deficiente, sin embargo, la palabra vacía no.

Para solucionar el problema de deficiencia, en IBM-5 (Brown et al. 1988) el modelo posiciona palabras únicamente en lugares disponibles. El modelo de distorsión toma en cuenta los lugares disponibles, usando un arreglo $[1; j]$ de la salida.

$$\text{Primera palabra del septo: } d_1(v_j|\mathcal{B}(e_j), v\odot_{i-1}, v_{\text{máx}}) \tag{2.14}$$

$$\text{Siguientes palabras: } d_{>1}(v_j - v_{\pi,k-1}|\mathcal{B}(e_j), v_{\text{máx}}) \tag{2.15}$$

Se conserva el reordenamiento relativo al septo del modelo 4, pero esta vez se limita el número de resultados generados por medio de $v_{\text{máx}}$. Se tiene también $v_{\pi,k-1}$ como el número de espacios disponibles en la palabra destino previa y, por lo tanto, $v_j - v_{\pi,k-1}$ es el número de espacios no evitados más uno. Esto es entonces un proceso de alineado de cada palabra, uno por uno.

El Modelo de Lenguaje (LM)

Para generar una buena traducción es necesario contar con un buen modelo del lenguaje, que servirá para presentar la traducción como un texto entendíble y legible en el idioma destino. Esto garantizará un orden de palabras correcto. La probabilidad $Pr(e)$ presentada en la ecuación 2.4 es planteada como un modelo de idioma objetivo e y se denota como p_{LM}. Se toma una sentencia en el idioma destino y regresa la probabilidad que ésta corresponda a una frase en ese idioma. Una frase correcta generará una mayor probabilidad que una incorrecta. Esta función permitirá al sistema traductor encontrar el orden correcto para las traducciones.

La implementación del estado del arte para p_{LM} son los modelos de n-gramas y se basan en la probabilidad de que una palabra siga después de otras antecesoras. Se desea

computar la probabilidad de una cadena $W = w_1, w_2, \ldots, w_n$. El problema surge cuando se quiere computar $p(W)$. Esto no tiene sentido, al generarse muchos datos dispersos. Por ello se recurre a la cadena de Markov. En vez de calcular $p(W)$, se va a calcular la probabilidad de una palabra a la vez. Descomponemos la probabilidad(Koehn 2010):

$$p(w_1, w_2, \ldots, w_n) = p(w_1)p(w_2|w_1)\ldots p(w_n)p(w_n|w_1, w_2, \ldots, w_{n-1}). \tag{2.16}$$

Como podemos observar la probabilidad $p(w_1, w_2, \ldots, w_n)$ es un producto de probabilidades de palabras dado un historial de palabras que le preceden. Se puede limitar el historial de palabras a m, con lo cual

$$p(w_1, w_2, \ldots, w_n) \simeq p(w_n|w_{n-m}, \ldots, w_{n-1}). \tag{2.17}$$

La secuencia de palabras sobre la cual se transita de una palabra a otra tiene, por lo tanto, transiciones con una historia limitada. Esta simplificación es llamada cadenas de Markov. El orden del modelo es el número m de palabras usadas como historial de una probabilidad. La base de n-gramas es una cadena de Markov con n de historia.

Para su estimación de un bigrama, se calcula la probabilidad de una palabra, dada dos palabras anteriores a ella.

$$p(w_2|w_1) = \frac{\text{count}(w_1, w_2)}{\Sigma_w \text{count}(w_1, w_2)} \tag{2.18}$$

2.4. Traducción por frases

En el apartado de alineación del modelo por palabras, se ha expuesto el problema de la monotonía de alineación. Esto es, el supuesto de que una palabra en la frase origen corresponde en el proceso de traducción, forzosamente, a una palabra en el idioma destino. Lo anterior es una afirmación que no se cumple en gran parte de los casos de traducción. Para resolver este problema, los modelos IBM 3,4 y 5 presentan la propuesta de integrar un modelo de fertilidad y distorsión, y así poder expresar una relación de uno-a-muchos. La relación se da desde f a e, sin embargo, no es posible expresar la relación varios-a-varios, limitándose a modelos con palabras.

En 1999 Franz Och, Christoph Tillmann y Hermann Ney plantean que una "apli-

cación más sistemática es considerar la frase completa, en vez de únicamente palabras individuales, como la base de los modelos de alineamiento"(Och et al. 1999). Por lo tanto, concluyen, que los cambios en los contextos de las palabras pueden ser aprendidos para influenciar también el orden de las palabras de salida. Su propuesta es la de creación de plantillas, que describen el alineamiento entre clases de secuencias, y una secuencia de clases de salida. Se continuó trabajando sobre el tema, hasta que Philipp Koehn, Franz Och y Daniel Marcu (Koehn et al. 2003), en 2003, presentan la base del modelo por frases.

Como ya se ha expuesto anteriormente, tenemos dos enunciados, uno origen f y uno destino e, donde $p(e|f)$ es la probabilidad de obtener una frase destino a partir de una origen. Usando el modelo del canal ruidoso y usando la regla de bayes obtenemos $\text{argmax}_e p(e|f) = \text{argmax}_e p(f|e)p(e)$.

Ahora, se divide f en I frases $\bar{f}_1^I$, y cada frase $\bar{f}_i \in \bar{f}_1^I$ es traducida en una frase e_i (Koehn et al. 2003). Esta traducción será modelada por la distribución $\phi(\bar{f}_i|\bar{e}_i)$. Como se toma en cuenta un reordenamiento de la salida en la frase destino se tendrá un modelo de distorsión relativa definido como $d(a_i - b_{i-1})$. La posición de inicio de la frase origen está denotada por a_i, y es traducida a la ́ésima frase destino. La posición final de la frase origen traducida a la $(i-1)$ frase destino se denota por b_{i-1}.

Se agrega también un factor ω para cada palabra destino generada, además del modelo de lenguaje p_{LM}. Para expresar lo anterior, se modifica el modelo de máxima esperanza presentado en la ecuación 2.4 de la siguiente manera:

$$e_{mejor} = \text{argmax}_e p(e|f) \tag{2.19}$$

$$= \text{argmax}_e p(f|g) p_{LM}(e) \omega^{\text{tamañoe}}. \tag{2.20}$$

El termino $p(f'^J|e^I)$ se descompone de la siguiente manera:

$$p(\hat{g}_1^I|\hat{s}_1^I) = \prod_{i=1}^{I} \phi(\hat{g}_i|\hat{s}_i) d(\text{inicio}_i - \text{fin}_{i-1} - 1) p_{LM}(e). \tag{2.21}$$

Con lo anterior, obtenemos un esquema de traducción con tres elementos: el modelo de traducción, el modelo de alineamiento o distorsión y el modelo del lenguaje destino. Esto lo podemos ver en la figura 2.3, al igual que la relación entre el entrenamiento y la decodificación.

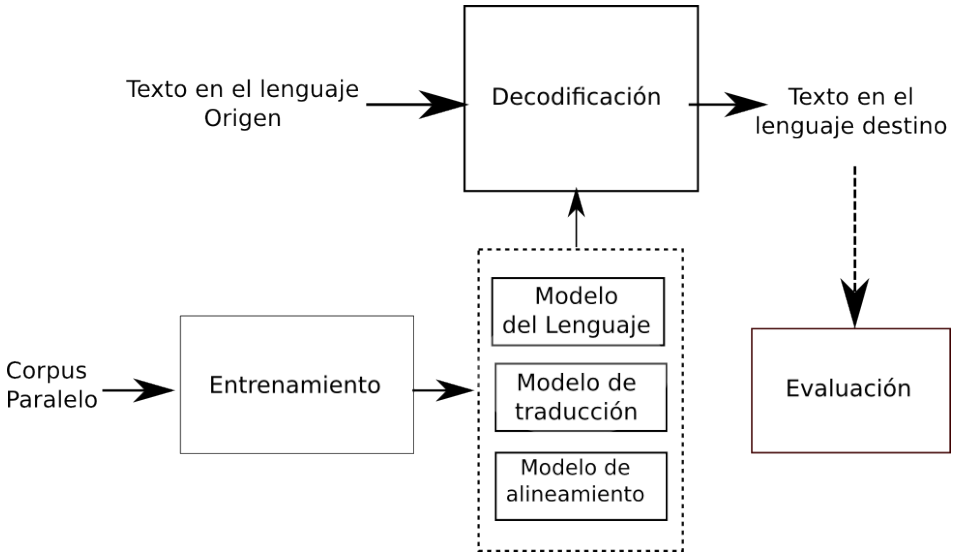

Figura 2.3: Traducción basada en frases

El modelo d es basado en distancias, que considera la nueva posición en la frase destino, relativa a la posición de la palabra en la frase origen. La distancia del reordenamiento es el número de palabras que se salta. La probabilidad de d se calcula con una función exponencial de decaimiento $d(x) = \alpha^{|x|}$, donde $\alpha \in [0, 1]$. Se aplica este criterio exponencial para imponer un costo mayor a las distancias grandes.

2.4.1. Entrenamiento

El proceso de entrenamiento requiere dos pasos previos, el de alineamiento y de entrenamiento de un modelo de lenguaje, para pasar posteriormente entrenar el modelo de traducción por frases. Con el entrenamiento del alineamiento, también se generan alineamientos de palabras de los datos de entrada (como por ejemplo con GIZA++ (Och & Ney 2003)). Con estos datos es posible colectar los pares de frases que son consistentes con el alineamiento, donde las palabras en los pares de frases legales únicamente se alinean entre ellas, y no con alguna otra (Och et al. 1999).

"Existe consistencia para el par de frases $(\bar{f}, \bar{e})$ y el alineamiento A, si todas las palabras $d_1, \ldots, f_n \in \bar{f}$, que tienen un punto de alineamiento en A, tienen estas palabras en $e_1, \ldots, e_n \in \bar{f}$ y viceversa"(Koehn 2010). Planteando esto de manera formal, obtenemos:

$$
\begin{aligned}
(\bar{e}, \bar{f}) \text{ son consistentes con A} \Leftrightarrow \\
\forall e_i \in \bar{e} : (e_i, f_j) \in A \Rightarrow f_j \in \bar{f} \\
\wedge \forall f_j \in \bar{f} : (e_i, f_j) \in A \Rightarrow e_i \in \bar{e} \\
\wedge \exists e_i \in \bar{e}, f_j \in \bar{f} : (e_i, f_j) \in A
\end{aligned}
$$

Ahora se puede pasar a la extracción de frases consistentes. La idea es iterar sobre todas las frases destino posibles, encontrando la frase mínima que sea consistente. En el algoritmo 1 (Koehn 2010), se muestra como se lleva acabo la extracción de frases.

A continuación será necesario calcular la tabla de probabilidades de traducción por frases. Una vez extraídos los pares de sentencias, se extrae el número de frases pares. Posteriormente se hace un conteo, en cuantos pares de sentencia es extraído un par particular. Este par es contabilizado mediante la función $count(\bar{e}|\bar{f})$, y con esto ϕ es estimado mediante:

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}|\bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e}, \bar{f}_i)} \tag{2.22}$$

Para agregar más de una frase, que tiene correspondencia con varias otras frases, se pueden asignar valores fraccionarios. Otro problema, que se enfrenta, es un crecimiento del tamaño de la tabla de traducción con respecto al del corpus. Si las tablas ocupan varios gigabytes, cargarlas a la memoria resulta en una dificultad para el equipo donde se ejecuta el proceso. En detrimento con la velocidad es posible usar memorias *swap* grandes o leer directamente las tablas desde el disco duro. Lo recomendable es, sin embargo, la utilización de equipos con una gran cantidad de memoria, normalmente varios cientos de gigabytes, para almacenar todas las tablas en la memoria y mejorar su velocidad de acceso. Para cargar las tablas de probabilidades por frases entrenadas con el el corpus Europal inglés - alemán, se requirieron ciento cuarenta gigabytes de memoria.

2.4.2. Decodificación

Con el modelo de lenguaje, el modelo de alineamiento y el modelo de traducción se busca la traducción con el mejor puntaje en el modelo expresado en la fórmula 2.6.

Algoritmo 1 Algoritmo de extracción de frases

entrada: alineamiento de palabras A para el par (e, f)
salida: conjunto de frases pares BP
para todo $e_{inicio} = 1, \ldots, |e|$ **hacer**
  para todo $e_{fin} = e_{inicio}, \ldots, |e|$ **hacer**
    para todo (e, f) **hacer**
      si $e_{inicio} \leq e \leq e_{fin}$ **entonces**
        $f_{inicio} \leftarrow \text{mín}(f, f_{inicio})$
        $f_{fin} \leftarrow \text{mín}(f, f_{fin})$
      fin si
    fin para
    $BP \leftarrow$ EXTRAER$(f_{inicio}, f_{fin}, e_{inicio}, e_{fin})$
  fin para
fin para
function EXTRAER$(f_{inicio}, f_{fin}, e_{inicio}, e_{fin})$
  si $f_{fin} = 0$ **entonces devolver** {}
  fin si
  para todo $(e, f) \in A$ **hacer**
    si $f_{inicio} \leq f \leq f_{fin} \vee (e < e_{inicio} \vee e > e_{fin})$ **entonces devolver** {}
    fin si
  fin para
  $E = \{\}$
  $f_s = f_{fin}$
  repetir
    repetir
      Agregar el par $(e_{inicio} \ldots e_{fin}, f_s \ldots f_e)$ al conjunto E
      $f_e = f_e + 1$
    hasta que f_e
    $f_s = f_s - 1$
  hasta que f_s esté alineado
  devolver E
fin function

Al ser este un problema NP-Completo(Knight 1999), se requiere el uso de heurísticos para encontrar una traducción aproximada. Se utilizan algoritmos como Beam o A^* (P. E. Hart & Raphael 1968) para ese fin. La búsqueda resultante es expresada en grafos, como se precia en la figura 2.5.

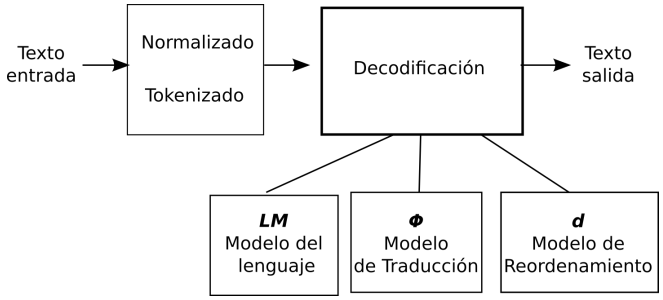

Figura 2.4: Proceso de decodificación.

En 2004, Philipp Koehn (Koehn 2004) desarrolla del traductor *Pharaoh*, que es un decodificador *Beam Serach* para SMT basada en frases. Su trabajo es la base para Moses, desarrollado por Hoang y Koehn (Hoang & Koehn 2008), el primer decodificador para el modelo de traducción por frases de código completamente libre, usando licencia GPL. Alignment Template System (ATS) (Bender et al. 2004), decodificador que usa A^* y traducción por plantilla, nunca fue distribuido públicamente, y *Pharaoh* únicamente fue distribuido de forma binaria, lo cual llevó a que se limitara el estudio del SMT y evitaba modificarlo para futuras experimentaciones. *Moses* ha sido desde entonces la alternativa para el estudio y la experimentación, y es el decodificador que se usa para este trabajo. El algoritmo *beam search* aplicado para la decodificación por frases tiene sus primeros estudios en la tesis doctoral de Tillmann en 2001 (Tillmann 2001) y Och 2002 (Och & Ney 2002), y fue ampliado e implementado por Koehn en *Pharaoh*.

Dada una tupla de palabras del lenguaje meta, este tiene una posibilidad de traducción para la frase original y es llamada opción de traducción. El algoritmo comienza con una hipótesis inicial, traduciendo una palabra de la frase. A partir de esa primer hipótesis, se expande una nueva hipótesis traduciendo una nueva palabra. Supongamos una traducción entre el español y el inglés. La frase *quiero ir hoy a comer comida china* iniciaría su traducción con una hipótesis inicial. Esta hipótesis puede ser *food* para comida, o *chinese* para comida. Al momento de expandir la hipótesis de *food* agregamos la traducción de alguna otra palabra, como *want* para querer. La hipótesis extendida puede volverse a extender, y así consecutivamente para cada caso. Las combinaciones resultantes de este espacio de búsqueda son exponenciales. En la figura 2.5 se muestra el

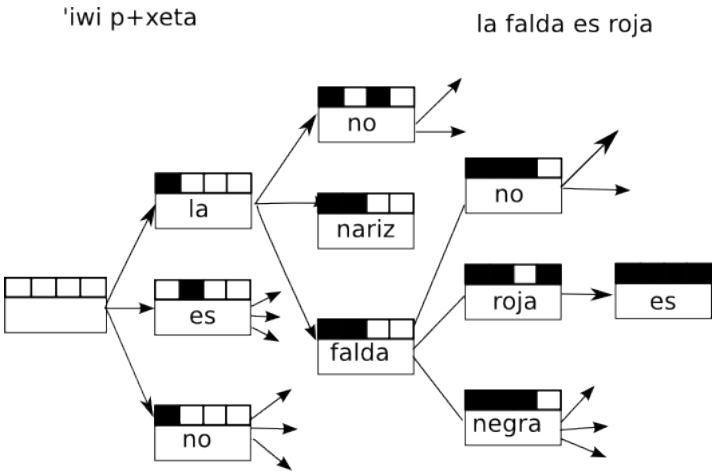

Figura 2.5: Búsqueda en el espacio por una traducción óptima

ejemplo para la frase wixárika *'wi p+xeta* y su expansión en un grafo de búsqueda. Dada la complejidad de la búsqueda se podará el espacio para hacer posible una búsqueda eficiente. Para realizar esta poda se recombinan hipótesis, bajo el siguiente criterio: dos hipótesis que contengan las mismas palabras; las últimas dos palabras generadas son iguales; y el final de las últimas frases cubiertas. Si existen dos hipótesis que coincidan en estas propiedades, se conserva únicamente la de menor costo(Koehn 2004).

Pero a pesar de la recombianción de hipótesis, el espacio sigue siendo de dimensión exponencial en su cota superior, siendo $A \simeq 2^{n_f}|V_e|^2 n_f$ (Koehn 2004), haciendo que el problema de la decodificación sea NP-completo. Es por eso que en el algoritmo Beam search se plantea la poda de las opciones de traducción inferiores, para todas las hipótesis que comparten el mísmo número de palabras traducidas.En el algoritmo 2, se muestra como funciona Beam Search. Cada pila descrita en el algoritmo corresponde al número de palabras destino traducidas. En el caso de la primer pila, se almacenará todas las hipótesis iniciales que contengan una palabra, mientras que la segunda pila contendrá las hipótesis de dos palabras y así sucesivmaente.

La hipótesis inicial se plantean con la traducción de las palabras más fáciles de traducir. Por lo tanto, el algoritmo no únicamente incluye el costo, como criterio de poda, sino la estimación del costo futuro. El tamaño de la pila se puede determinar por una memoria de portal. Con este portal definido como α se descarta toda aquella probabilidad inferior al mismo.

Algoritmo 2 Heurística "Stack Decoding"

```
Agregar una hipótesis vacía a la pila 0
para todo pilas 0, ..., n − 1 hacer
    para todo hipótesis en la pila hacer
        para todo opciones de traducción hacer
            si es aplicable entonces
                h ← Crear una nueva hipótesis
                agregar h a la pila
                Recombinar con hipótesis existentes si es posible
                Podar la pila si es posible
            fin si
        fin para
    fin para
fin para
```

2.4.3. Los modelos híbridos y los retos para el caso particular de la traducción con bajos recursos

Si bien el trabajo expuesto anteriormente presenta una metodología apta para traducción automática usando gran cantidad de datos, el problema de pares de lenguas con pocos datos es real. La SMT depende enteramente de la cantidad de datos con que se cuenta en el corpus paralelo. Para el uso de modelos de traducción estadística del estado del arte serían necesarios 100 megabytes de texto pre-alineado (Laukaitis & Vasilecas 2007), lo cual con idiomas como el wixárika sería imposible de obtener en este momento. En 2002 se plantea, por primera vez, el problema de SMT en lenguajes con pocos recursos paralelos (Al-Onaizan et al. 2002*a*). El caso del traductor wixárika-español comparte el problema antes descrito. Para trabajar este problema se ha propuesto utilizar modelos híbridos cómo Laukaitis (Laukaitis & Vasilecas 2007), Yaser (Al-Onaizan et al. 2002*b*) y Nießen (Nießen & Ney 2004). Asumir una traducción gramatical, basada en reglas, tampoco es posible por la falta de un cuerpo completo de la gramática wixárika, a pesar de los avances en los últimos años en la materia.

Nießn y Ney proponen la utilización de un analizador morfológico que descomponga las palabras en sus raíces y morfemas para etiquetar posteriormente cada componente. Se auxilia de un diccionario jerárquico. Este mecanismo logra reducir el corpus paralelo necesario hasta a 10 % del normalmente usado. Laukaitis (Laukaitis & Vasilecas 2007) analiza el caso de un traductor asimétrico, donde un lenguaje tiene una gran cantidad de recursos y el segundo carece casi por completo de ellos, con excepción de un analizador

morfológico. Con ayuda de un corpus paralelo reducido (de 0.2 megabytes) y redes ontológicas del lado del idioma más analizado, logra buenos resultados. Además, es posible utilizar técnicas de compiladores, o la creación de vectores de palabras relacionadas.

El analizador morfológico de los idiomas aglutinantes, como es el caso del wixárika, puede ser representado por medio de transductores de Estado Finito, cómo también ha sido realizado para el turco (Eryiğit & Adalı 2004). Ermolaeva (Ermolaeva 2014) también plantea un analizador morfológico, pero adaptativo para los lenguajes aglutinantes en general. Al ser el wixárika una lengua polisintética y contar con una gran variedad de palabras generadas a partir de los morfemas, no sería posible tomar los parámetros de medición por palabras, cómo en los modelos de lenguaje por palabras. La separación por morfemas, por el contrario, permitiría una mejor traducción.

Capítulo 3

Metodología

En este capítulo se presenta la metodología propuesta para el problema de traducción estadística de la lengua wixárika al español y viceversa, y de una manera más general de una lengua algutinante a una fusionante. Todo esto se realizará en el contexto de escasos recursos de traducción para su entrenamiento.

Para comenzar se van a presentar propiedades generales del idioma wixárika, con lo cual se presentan conocimientos gramaticales que posteriormente se aprovecharán en la metodología de traducción híbrida. Después se presentará la metodología utilizada y el trabajo morfológica que se emplea.

3.1. El idioma wixárika

El wixárika es un idioma perteneciente a la familia yutonahua como se muestra en la figura 3.1, con una estructura sujeto-objeto-verbo(SOV), incorporante y con una fuerte tendencia polisintética, que es incluso mayor que la del náhuatl. Los morfemas se agrupan en torno a una raíz verbal (ver figura 3.2) e incluyen una gran cantidad de información. La polisíntesis es el resultado de la incorporación de operaciones sintácticas, realizado en otros casos por la combinación de palabras autónomas a la palabra predicativa, aproximándose al ideal de una palabra por enunciado (Iturrio & Gómez López 1999).

La familia yutoazteca o yutonahua, cuenta con dos grandes divisiones. La primera, las lenguas septentrionales, las cuales se hablan principalmente en Estados Unidos y la segunda son las meridionales, en el norte de México. Varias de las lenguas septentrionales se encuentran extintas o en proceso de desaparecer, con excepción del hopi (5 mil

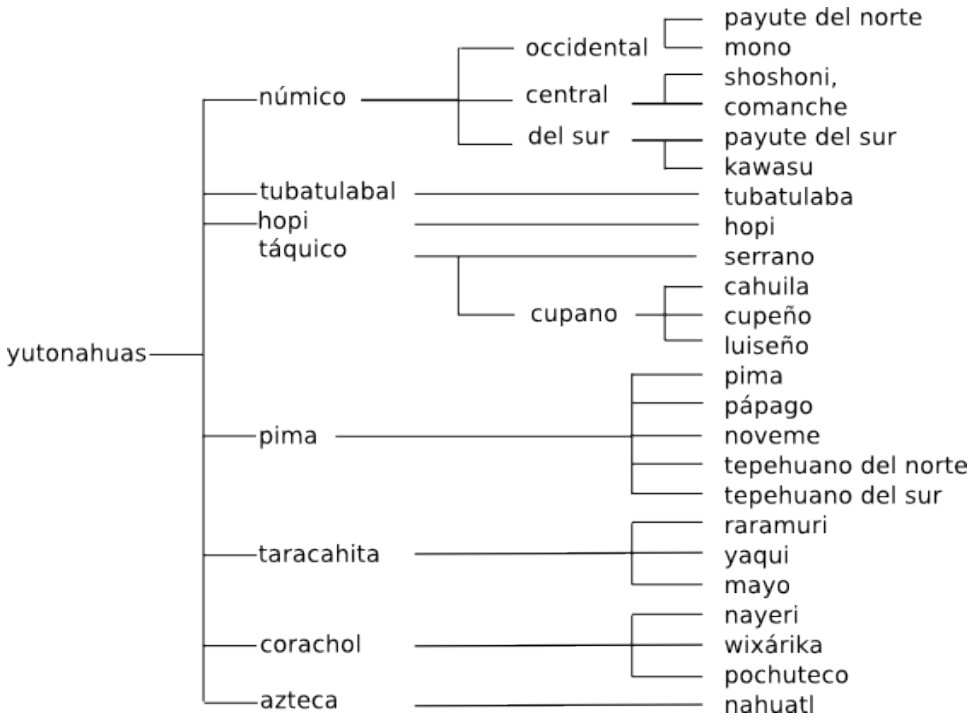

Figura 3.1: La familia de las lenguas yutonahuas tomada de(Iturrio & Gómez López 1999)

hablantes), el comanche, pauite y shoshoni (con menos de 5 mil hablantes). En México, las lenguas con más de mil hablantes son el náhuatl (1 millón 725 mil), o'dam (46 mil) yaqui (20 mil), mayo (42 mil), rarámuri (73 mil), náyeri (28 mil) y el wixárika (52 mil) según el INEGI (INALI 2016). El náhuatl se ha extendido en el centro y sur del país por las migraciones desde el norte en la época prehispánica. Todas estas lenguas comparten la estructura sintáctica y la característica aglutinante. Con la actual metodología, que se presenta, y que se basa en las características de la familia lingüística yutonahua, será posible, en un futuro expandir la implementación a las otras lenguas de la familia.

En la figura 3.2 se muestra el triángulo de Helwag, que nos explica la relación de las vocales para las lenguas europeas. En el caso de wixárika no existe la vocal *o*, pero se agrega una quinta vocal escrita como *+* y que se asemeja a la letra alemana *ü*.

El conjunto de símbolos Σ del wixárika se presenta en la tabla 3.1. También se muestran los símbolos usados para escribir el wixárika en otras variantes a las usadas en el presente trabajo. Algunos de estos símbolos se encuentran en desuso o se ocupan únicamente en el ámbito académico. La falta de unidad en la escritura wixárika dificulta su uso en el procesamiento de lenguaje natural. Por ejemplo, en Palafox (Vargas 1978) wixárika es escrito como huirrárika, o en otros wirrarika, mientras que en la actualidad

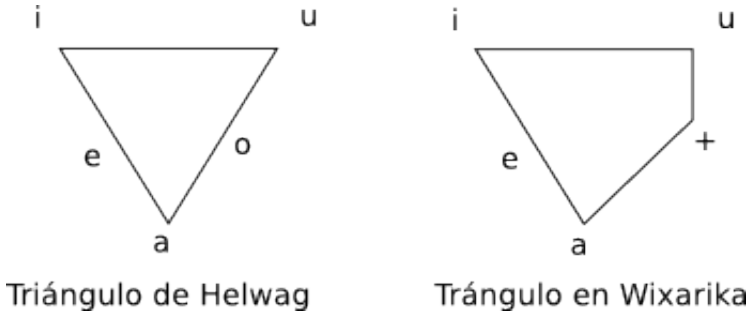

Figura 3.2: Triangulo de Helwag y Helwag modificado para el wixárika(Iturrio & Gómez López 1999)

se utiliza la palabra wixárika.

Con el conjunto de símbolos descritos se crean las palabras y frases del wixárika, sin embargo, los verbos y en menor medida los sustantivos, son formados por morfemas, que se aglutinan en torno a la raíz. En el siguiente ejemplo se aprecia la forma en que se pueden construir palabras en wixárika a partir de sus reglas silábicas. El concepto de montaña puede ser creado de la siguiente manera

hai m-a-ta-ka-i-t+ka

Donde *hai* significa nube y la palabra siguiente es el verbo *matakait+ka* que se divide en morfemas. La combinación entre *m* y *a* refiere a algo figurativo, el *ta* a algo que está al borde de, *ka* localiza esto en cierto espacio, la *i* significa estar mientras que *t+ka* es plural. El resultado puede ser leído como "donde las nubes bordean", y que en una manera muy generalizada se traduciría como montañas (Gómez 1999).

Como se ha mostrado, los verbos o sustantivos son formados por una serie de morfemas, los cuales pueden aglutinarse antes de la raíz de la palabra, o después de esta. En la tabla 3.2 se muestran los prefijos, mientras que en la tabla 3.3 se pueden ver los postfijos. Cada uno de estos morfemas varía o agrega significados a la palabra, y cada morfema tiene un uso diferente según la posición que ocupa. De tal suerte el prefijo *ka* tiene dos significados. Si se encuentra en la posición 16, se niega el enunciado; mientras que si se encuentra en la posición 3 o 1 caracteriza un movimiento hacia abajo.

Símbolo usado	Otras formas de expresar el símbolo
a	
e	
h	
i	
+	ü, Λ, ï, cu, ɨ
k	
kw	
m	
n	
p	
r	
t	
ts	tz, ch,
u	
w	v
x	rr,
y	
’	?, sin anotar

Tabla 3.1: Símbolos del wixárika

19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
ya ke ’a ’e ’u xa	’a tsi ke ’u	ne pe te xe me	ka	p+ m+ r+ka n+ka ke	ka	ka	ne ma ta xe	ti tsi	ta xe	ni	wa	’u ha he heu	’i	’ana ’anu wa	ne ’a ta yu	ti ta ku ka ye	red	ti ta ku ka ye	raíz

Tabla 3.2: Prefijos del verbo en wixárika

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
t+	ka y+ ya y+k+ m+	t+a ta ya rie kie ke ma wie xie	ka ta ya rie	tsi	t+a rie	ya	rie	x+a	xime r+me ne/ t+we we/ wawe m+k+/ ku mie/ yu n+a ka wa	t+ t+ka	kai/ tei x+ ni/ m+k+	x+a	ke	kai yu	tsie yari	k+ t+ kaku ka ku me yu	t+	ni ka	t+	ka(i)	t+	kaku

Tabla 3.3: Postfijos del verbo wixárika

3.2. Diseño y modelación del traductor

El modelo del traductor estadístico propuesto se compone de las implementaciones discutidas en la sección anterior, con el alineador GIZA++ que puede realizar las tareas utilizando cualquiera de los cinco algoritmos discutidos con mejoras, junto con el entrenador y decodificador por frases Moses. A estas herramientas del estado del arte, se preprocesa la entrada del corpus en wixárika. Para el presente trabajo se ha desarrollado un normalizador, tokenizador y un segmentador morfológico. Como se muestra en la figura 3.3 los nuevos elementos que se agregan al proceso de entrenamiento con los apartados del analizador, segmentador y etiquetado. Para el proceso de traducción wixárika a español también se agregan los elementos mencionados antes de la decodificación, como se muestra en la figura 3.4. En la traducción español a wixárika se agrega un aglutinador de los morfemas generados a partir de la decodificación 3.7. En los tres procesos mencionados se han añadido elementos que permiten la traducción de un idioma aglutinante.

En la fase de entrenamiento, a partir de un corpus paralelo entre el wixárika y el español generará tres modelos: el modelo del lenguaje p_{LM}, la tabla de traducción de frases $\phi(\bar{f}|\bar{e})$ y el modelo de alineamiento d, descritos en el capítulo anterior. Para el wixárika será necesario agregar una etapa de análisis y descomposición morfológica antes de pasar a la decodificación en la traducción wixárika al español, y un reordenamiento morfológico y aglutinación de los morfemas en la traducción del wixárika al español. En la fase de entrenamiento que se muestra en la figura 3.3 también se requerirá de un analizador y de un segmentador morfológico del corpus introducido. Esto separará las palabras wixárika en sus componentes morfológicos, los cuales serán tratados como equivalentes a palabras en los modelos SMT descritos en el estado del arte.

Para la evaluación y retroalimentación, se utiliza TER, WER y BLEU, además de una evaluación humana incorporada en plataforma web, donde los usuarios pueden agregar traducciones correctas o evaluar las traducciones generadas por el sistema.

3.2.1. Proceso de entrenamiento

El proceso de entrenamiento requiere un corpus paralelo, con dos archivos que contengan frases emparejadas (f, e), tal que puedan ser analizadas como equivalentes. Los datos presentados no tienen el supuesto de estar normalizados por lo que no podrán ser alineados inmediatamente.

Las frases en wixárika tienen un problema común. No existe una escritura unificada del mismo entre los hablantes a la hora de comunicarse por texto. Extenso ha sido el debate entre diferentes académicos sobre cuál debe ser Σ para el wixárika, sin embargo, esto no se ha logrado por completo. La SEP, en sus libros de texto ha logrado impulsar un cierto acuerdo entre los hablantes, el cual se usa de manera variada en los textos cotidianos wixaritari. También la ortografía varía y la misma unión de morfemas, en ocasiones se encuentra de manera irregular. Esto agrega un problema de normalización. Para el presente trabajo se ha creado un normalizador con expresiones regulares, que suprime el uso de acentos, dada la variación en usos del mismo, en diferentes textos, así como el cambio símbolos equivalentes a nuestro Σ. También se evita el uso de letras largas, como "aa" a "a", por diferencias en escrituras. Si bien, estos dos factores aportan información relevante sobre el sentido, al no existir un uso homogéneo de su escritura, confundirá al sistema. En la tabla 3.4 se muestra la normalización que utiliza el sistema. Cabe destacar que el símbolo usado por gran parte de los textos oficiales modernos para expresar la vocal + es ɨ. Un problema fuerte al usar este símbolo es su dependencia a un texto con formato. Para su uso en las herramientas de lenguaje natural esto no es posible. También en el uso cotidiano del wixárika se evita esta forma, por ser imposible de utilizar en las redes sociales, chats, etc. Es por ello que el símbolo + es más apropiado.

Una vez normalizado el texto se pasa a tokenizarlo. Este proceso se refiere a definir que es una palabra y como es posible distinguirla. En el presente caso se utiliza el carácter de espacio para diferentes palabras dentro de una frase, y a los signos especiales (punto, interrogación, exclamación, etc.) también se consideran como palabras. La expresión regular para realizar esta tarea se presenta en la tabla 3.5

Con estas anotaciones sobre el normalizado y el tokenizado, se puede pasar al analizado morfológico. Es una gran ventaja contar con un corpus previamente segmentado por humanos, como es nuestro caso. Este corpus ha sido extraído de Gómez (Gómez 1999) y se ha transcrito de la siguiente forma.

```
wani p+-ne-tsi-'u-ti nanai-t+a=juan me hizo reír
mexa ne-p-eu-xawa-ri-ya-x+=agujereé la tabla
mexa p-eu-xawa=la tabla está agujereda
tsik+iwiti ne-p-u-ta-haxu-ma=enlodé la canasta
```

Con esto es posible separar los morfemas, y anotar la posición que ocupan dentro de la tabla morfológica 3.2 y 3.3. A continuación se muestra la forma en que se anotan

Operación	Cadena original	Cadena resultante
Texto a minúsculas		
Sustituir	´	’
Sustituir	v	w
Sustituir	c	k
Sustituir	[0-9]+	vacío
Sustituir	ch	ts
Sustituir	rr	x
Sustituir	espacio+	espacio
Sustituir	[áàä]	a
Sustituir	[éèë]	e
Sustituir	[íì]	i
Sustituir	[óòö]	o
Sustituir	[úù]	u
Sustituir	[ïü]	+
Sustituir	([a-z])\1+	\1

Tabla 3.4: Normalizador del wixárika

Operación	Cadena original	Cadena resultante
Sustituir	[^\s]($[.|,|,\-,\"|:|;|¿|?|¡|!]$)	espacio\1
Sustituir	[.|,|,\-,\"|:|;|¿|?|¡|!])[^\s]	\1espacio

Tabla 3.5: Tokenizador del wixárika

las posiciones de cada morfema, para el entrenamiento. El mismo resultado puede ser generado a partir del analizador morfológico, con un grado de error.

```
wani p+ ne tsi 'u ti nanai t+a-19
mexa ne1 p3 eu11 xawa
mexa p3 eu11 xawa
tsik+iwiti ne1 p3 u11 ta14 haxu ma-22
```

La separación del texto alineado en dos archivos, su normalización, tokenización y su anotación morfológica se realizan de manera automatizada mediante un *script*. El mismo llama a las siguientes etapas del entrenamiento. Se ha incorporado un diccionario para el trabajo de análisis morfológico, donde las palabras que no deben ser segmentadas por poseer un cuerpo no aglutinado, se encuentran en un archivo. Estas palabras, a su vez, contienen una relación directa con palabras en español, por lo que son utilizadas para ayudar a enriquecer el corpus o para mejorar el alineamiento de palabras.

Una vez realizadas las tareas antes descritas, se pasará al alineado de palabras, a la obtención de la tabla de traducción léxica y al entrenamiento del modelo del español y del wixárika. En el caso del modelo del wixárika será necesario entrenar el modelo con un corpus ya descompuesto morfológicamente, para poder calcular las probabilidades correctas en la fase de traducción. Con los tres modelos entrenados se podrá pasar al cálculo de las tablas de traducción ϕ, cuyo proceso consta de extraer frases que sean consistentes, evaluar las frases y construir un modelo de reordenamiento, tal como se muestra en la figura 3.3, junto con el conjunto de la metodología de entrenamiento. En esta figura también se introduce la nueva fase morfológica que permite reducir el corpus necesario para la traducción y mejorarla.

Con fines de comparación, se ha entrenado el traductor con tres formas distintas. La primera consiste en tener un traductor sin modificar el proceso de traducción de Moses, con el fin de tener un referente de traducción del estado del arte con el cual se compara la presente metodología. El segundo entrenamiento se realiza con segmentación morfológica, pero sin etiquetado de cada morfema según su posición y, por último se entrena con segmentación morfológica y con anotación de posición para cada morfema.

3.2.2. Proceso de traducción

Con los tres modelos necesarios, es posible comenzar con la traducción. Se tendrán dos casos: el primero a tratar será el del wixárika al español, mientras que el segundo es del español al wixárika.

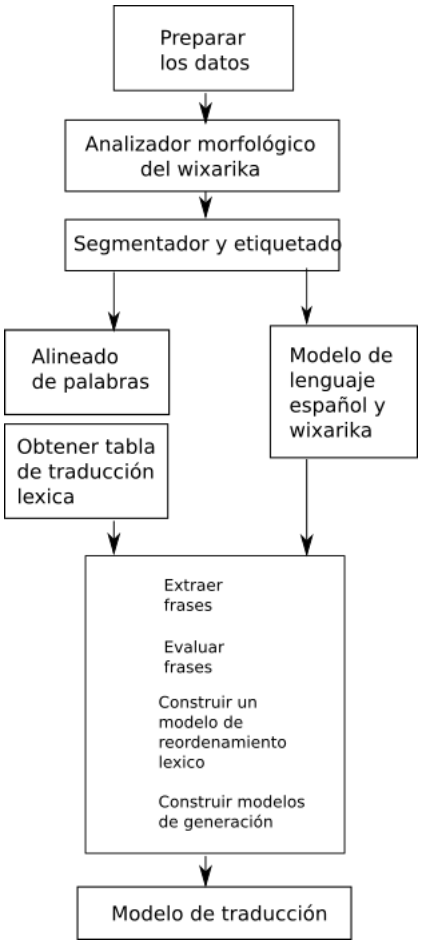

Figura 3.3: Proceso de entrenamiento

3.3. Proceso de traducción

Para la traducción wixárika a español, será necesario utilizar nuevamente el normalizador propio del wixárika (tabla 3.2) en el texto entrante, para poder mantener un alfabeto común con el corpus. También se usará el tokenizador (table 3.3) creado para este lenguaje. Una vez preparado el texto, se podrá pasar a la fase de análisis morfológico y segmentación. En este caso, es importante distinguir entre una palabra que debe ser segmentada y otra que existe en un diccionario de palabras, las cuales no responden a una lógica aglutinante. Dado que se cuenta con un diccionario, y en este se encuentran palabras no aglutinadas, se recorrerá toda la frase, y todas aquellas palabras que existen en el diccionario serán conservadas íntegras, mientras que las desconocidas serán analizadas por la función Ξ. Las palabras serán integradas nuevamente y presentadas al decodificador, tal como se muestra en la figura 3.4, el cual aproximará una máxima

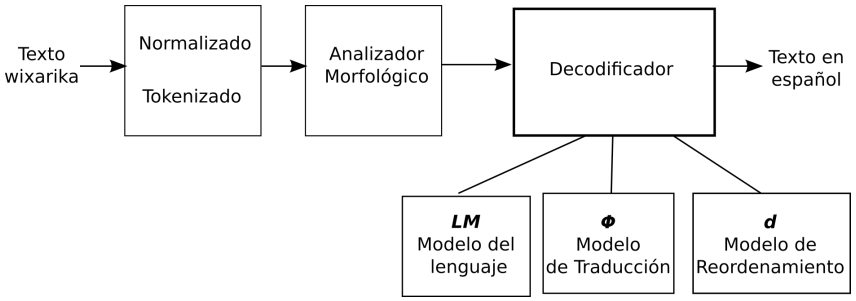

Figura 3.4: Proceso de decodificación del wixárika al español.

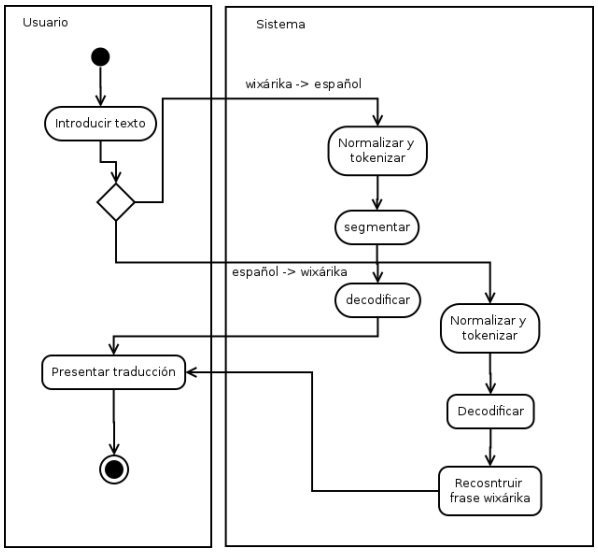

Figura 3.5: Diagrama de actividades (traductor)

esperanza de traducción, mediante *Beam Search* 3.6 .

En caso de ser una traducción español a wixárika no se usará el analizador morfológico. Las palabras del español se agregarán íntegramente al modelo, usando un normalizado y tokenizado estándar para el idioma español, integrados a Moses. Una vez el texto se encuentre listo, será introducido al decodificador, lo cual generará una salida de morfemas y palabras. El corpus con el cual se ha entrenado el modelo del lenguaje wixárika no es extenso y, por lo tanto, el ordenamiento de los morfemas es complicado usando únicamente el LM. Para esto se ha utilizado un ordenador de morfemas, que se presenta en el algoritmo ...

Dado que tanto el entrenamiento como la decodificación requieren de un tratamiento morfológico, se pasará a presentar el planteamiento del mismo y los algoritmos que se han desarrollado para realizar esta tarea.

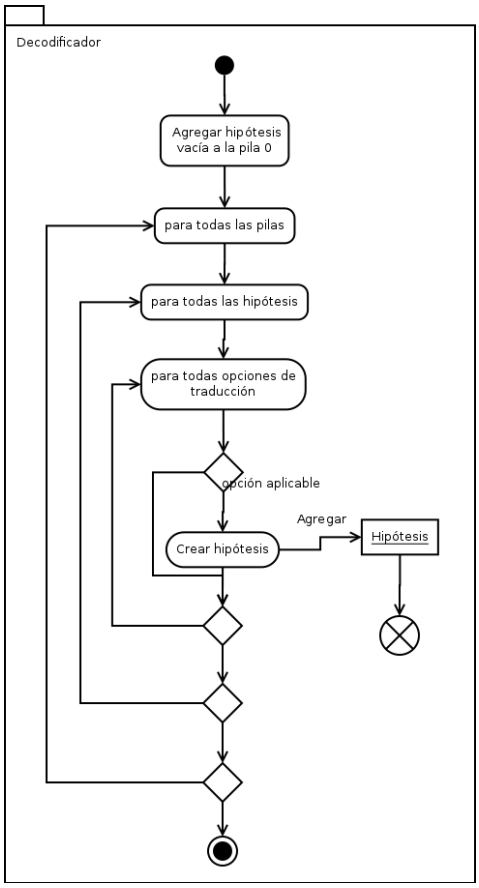

Figura 3.6: Decodificador

3.4. Tratamiento morfológico

Dado que el wixárika es un lenguaje polisintético, el alineamiento con palabras al español es poco prometedor. Los afijos se aglutinan en torno al verbo y al sustantivo, tanto antes de la raíz como después. La función $j \rightarrow i = a_j$ no se cumple como un mapeo uno a uno, como lo sugieren los modelos IBM 1 y 2, sino en forma de $j \rightarrow (i_1 \dots i_n) = a_j$ donde $k \geq 1$ y a_j es una tupla de pares de alineamiento, que ya conocemos como la fertilidad ϕ, de los modelos IBM 3, 4 y 5. Sin embargo, a pesar de que estos modelos toman en cuenta la fertilidad de una palabra del wixárika con respecto a varias del español, se requiere que cada palabra wixárika, se alinea a frases completas del español, como se muestra en la figura 3.8. Esto requeriría de un gran corpus, donde se necesita por lo menos una aparición de cada combinación posible de morfemas para cada raíz. Pero tomando en cuenta la reducida cantidad frases alineadas, un modelo de fertilidad sin mayor cambio, no sería posible.

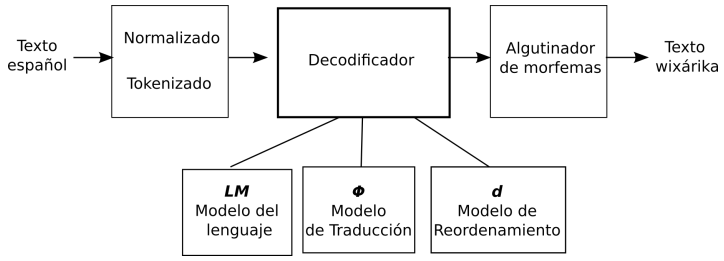

Figura 3.7: Proceso de decodificación del español al wixárika.

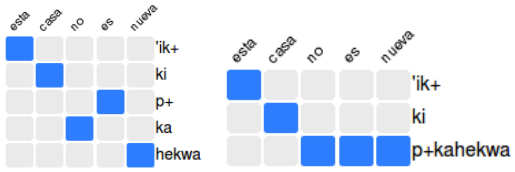

Figura 3.8: Búsqueda de la mejor traducción

Para el caso de idiomas con gran riqueza morfológica, como el caso de estudio, se sugiere separar los morfemas que sean más parecidos a palabras del inglés, conservar unidos los morfemas (como tiempos verbales) a sus raíces, que se comporten de manera semejante en inglés, e ignorar los que no tienen funciones parecidas (Koehn 2010). Se va a tomar la concatenación de todas las palabras y morfemas generados por el algoritmo 3 para conservar la mayor cantidad de información posible.

Ejemplificamos este alineamiento con una frase en wixárika. El alineamiento se realiza de la siguiente manera. Se tiene dos frases: “esta casa no es vieja” y “ik+ ki p+ ka ’ukiratsi”. La frase en wixárika ha sido descompuesta morfológicamente. A continuación se muestra la forma en que se alinean las dos frases.

```
esta casa no es vieja
NULL ({ }) ’ik+ ({ 1 }) ki ({ 2 }) p+ ({ 4 }) ka ({ 3 }) ’ukiratsi ({ 5 })
```

3.4.1. Propuesta

Para poder realizar la segmentación se tendrá una función Ξ que permitirá descomponer una palabra en sus morfemas. Esta función es un transductor de estados finitos, que tendrá como auxiliares las tablas 3.2 y 3.3, un diccionario de palabras que no deberán ser segmentadas y un lista de raíces verbales. Tanto el diccionario, como la lista de raíces puede ser introducida manualmente o aprendida de un corpus segmentado.

Como se muestra en el algoritmo 3, si una palabra se encuentra en el diccionario

D entonces esta palabra se agregará sin modificaciones a la cadena de salida W, pero si no aparece será evaluada por Ξ y el resultado agregado a la salida W. Esta salida es una tupla que contiene los morfemas que componen la palabra; si no se encuentra segmentación acorde a las tablas morfológicas, entonces se devolverá la palabra original sin cambios.

En el modelo de traducción, en vez de usar la tupla f se usa el $f^m = \Xi(f)$, el cual se sustituye en la ecuación 2.20. La figura 3.8 muestra la mejora en el alineamiento de palabras del modelo de descomposición morfológica al modelo de alineamiento de palabras. La frase *ik+ ki p+kahekwa* se traduce como *esta casa no es nueva*. Pero la palabra *p+kahekwa* contiene la información de tres palabras en español. Si usamos nuestra función Ξ(p+kajekwa) obtendríamos la tupla $(p+, ka, hekwa)$. La unión de todas las palabras descompuestas y no descompuestas de la frase original f^I nos genera un mejor alineamiento respecto al español. La función Ξ es un transductor de estados finitos, con la información morfológica descrita en (Iturrio & Gómez López 1999) y (Gómez 1999). Los idiomas polisintéticos y aglutinantes comparten la característica de poder ser expresados mediante un transductor de estados finitos, como es el caso del turco (Eryiğit & Adalı 2004) (Ermolaeva 2014).

Algoritmo 3 Función Ξ

Entrada: línea de texto f, tabla hash diccionario D
Salida: Una lista ordenada W
function $\Xi(f)$
    Lista $W \leftarrow \emptyset$
    Lista $tokens \leftarrow$ DIVIDIR(f)
    para todo $token \in tokens$ **hacer**
        si $D[token] = \emptyset$ **entonces**
            AGREGAR(W, $token$)
        si no
            para todo $m \in \Xi(token)$ **hacer**
                AGREGAR(W, m)
            fin para
        fin si
    fin para
    devolver W
fin function

Usando GIZA++(Och & Ney 2000) observamos que los morfemas wixárikas corresponden en general con una palabra en español. Es posible que la correspondencia sea mayor a una o incluso que no exista relación. Con esta información se generará un modelo con el cual podemos reordenar la traducción por palabras en el lenguaje destino.

3.4.2. El segmentador

A continuación se presenta el algoritmo del segmentador morfológico propuesto, como un transductor de estados finitos (*Finite State Transducer* FST), para las funciones de reconocimiento y de traductor. Se define de manera formal un transductor según Jurafsky (Jurafsky & Martin 2000) de la siguiente forma:

- Q Conjunto de N estados $q_0, q_1, \ldots, q_{n_1}$
- Σ Un alfabeto de entrada
- Δ Un alfabeto de salida
- $q_0 \in Q$ El estado de inicio
- $F \subset Q$ El conjunto de estados de salida
- $\delta(q, w)$ La función de transición entre dos estados.
- $\sigma(q, w)$ La función de salida dado el conjunto de posibles cadenas de salida por cada estado de entrada.

Es un transductor de estados finitos, que utiliza tres arreglos de datos, con conocimientos gramaticales previos. El primer arreglo que se utilizará es un arreglo de todos los niveles de prefijos al verbo, definido por *pre*. En total son 17 lugares, cada uno con un número variable de morfemas posibles.

Primero se definen los estados de entrada Q como cada morfema y raíz y que se encuentran guardados en las listas r para las raíces y *pre* y *pos* para los prefijos y postfijos respectivamente, en los estados de entrada Q, tal que $Q = pre + r + pos$. Los estados de salida F son todos los estados de la lista *pos*. La función δ será de todo estado en el nivel i a todo nivel $n > i + 1$. Por lo que el avance en las transiciones será hacia niveles superiores, pero nunca hacia atrás o al mismo nivel donde se encuentra. Por último, las funciones de salida σ generarán una concatenación del estado actual y el

número de su nivel, un identificador si es raíz o un nivel negativo si es postfijo, junto con un espacio en blanco.

El algoritmo 4 llama a una función recursiva *start*, donde se introduce la posición, *pos*, en las tabla 3.2, que será cero en un inicio; una lista *path* que contendrá los morfemas recorridos; la palabra a analizar w; y una cadena que contendrá la cadena ya analizada. En la llamada inicial esta cadena estará vacía.

Algoritmo 4 Segmentador morfológico

```
Entrada: palabra w, arreglos r, pre y pos
START(w, "", 0, [])
function START(w, prev, pos, path)
    si pos > |pre| − 1 entonces devolver
    fin si
    gotone ← False
    para todo s ∈ pre[pos] hacer
        m ← Regex.match("^-prev+s+-")
        si m entonces
            gotone ← True
            npath ← path
            npath.append((pos, s))
            START(w, m, pos + 1, npath)
            para todo s ∈ r hacer
                sm ← Regex.match("^-nprev+s+-")
                si sm entonces
                    nspath ← npath
                    npath.append((0, sm))
                    END(w, sm, nspath)
                fin si
            fin para
        fin si
        si ¬gotone entonces
            si pos > 17 entonces devolver
            fin si
            START(w,prev, pos + 1, path)
        fin si
    fin para
fin function
```

Como primer acción, la función *start* probará si se ha alcanzado el total de posiciones posibles, que es la cardinalidad del arreglo *pre*, si se ha excedido el máximo número de posiciones; entonces la función retornará sin valor. La bandera *gotone* se

activará únicamente si la función logró encontrar algún morfema o raíz concatenado con la cadena ya analizada, pero de entrada es falsa. Una vez asignada la bandera inicia un ciclo que itera sobre todos los morfemas posibles que se encuentran en una posición de la tabla morfológica. Si existe, encuentra un morfema coincidente; entonces se activa la bandera *gotone*, se crea una lista que contenga el nuevo morfema agregado a la lista del *path* anterior, y se llama a si mismo con los nuevos valores y una posición superior. Si en el nivel existen más de dos coincidencias serán llamadas todas las coincidencias posibles. Si se encuentra, además de una coincidencia, una raíz posterior a la cadena evaluada, entonces se buscará raíces del diccionario de raíces que puedan coincidir. En dado caso de que se encuentren, se llamará a la función *end*, que será descrita en el algoritmo 5. Al finalizar la función, y una vez concluido los dos ciclos, si la bandera *gotone* se encuentra apagada, entonces se llamará a si mismo con los valores de entrada, pero con una posición incrementada.

Algoritmo 5 Segmentador morfológico 2

```
function END(w, prev, pos, path)
    si pos < 0 ∨ pos ≥ |post|) entonces devolver
    fin si
    si |prev| = |w| entonces
        PRINT(path)
    fin si
    gotone ← False
    para todo s ∈ post[−pos] hacer
        m ← Regex.match("^-prev+s+-")
        si m entonces
            gotone ← True
            npath ← path
            npath.append((pos, s))
            END(w, m, pos + 1, npath)
        fin si
        si ¬gotone entonces
            END(w, prev, pos + 1, path)
        fin si
    fin para
fin function
```

La función *end* será llamada hasta que se halla logrado encontrar una raíz, y por lo tanto se pasará a los post fijos. La condición de finalización será que $pos < 0 \vee pos \geq |post|$. La segunda condición es la terminación positiva, donde si el tamaño de la cadena

prev es igual al tamaño de la palabra a analizar, se considera que se ha concluido de analizar toda la palabra. Si no es el caso, seguirá analizando. Los caminos que no logran encontrar coincidencias con morfemas y raíces, para toda la cadena a analizar, serán descartados. La búsqueda de coincidencias se retomarán de la función *start*, al igual que la forma de avanzar si la bandera *gotone* se encuentra apagada.

Este método encuentra todas las posibles segmentaciones de una palabra, tomando en cuenta una morfología analizada a priori, y un diccionario de raíces. No se tiene un criterio para escoger entre varias segmentaciones, así como también se tiene la certeza de que si no se encuentra la raíz en el diccionario, entonces la segmentación puede realizarse erróneamente o no hacerse.

3.4.3. El diccionario y las raíces

Para poder realizar de manera correcta la segmentación se requiere un diccionario de palabras no aglutinadas y un diccionario de raíces. Si bien estos dos diccionarios pueden ser creados manualmente, también pueden ser extraídos de un texto segmentado, método que se ha usado en combinación con la recolección de un diccionario creado por humanos.

En el algoritmo 6 se encuentra las palabras no aglutinadas de un una línea de texto segmentado. Primero se dividirá la línea en palabras, cada palabra se recorrerá, y se intentará segmentar por el caracter "−". Si el tamaño del la tupla resultante es igual a uno, en otras palabras, no tiene segmentaciones, entonces se agregará al conjunto de palabras no aglutinadas. De lo contrario, se invocará al algoritmo 7.

Algoritmo 6 Encontrar palabras no aglutinadas

```
function GET(line)
    Set W ← {}
    para todo word ∈ line hacer
        w ← SPLIT(word)
        si |w| = 1 entonces
            ADD(W, w)
        si no
            SEGREG(word)
        fin si
    fin para
fin function
```

En el algoritmo 7 se aspira a dos cosas: anotar un texto segmentado con las posi-

ciones de cada morfema, según el nivel que ocupa, y encontrar todas las raíces posibles y guardarlas en un archivo, que será el diccionario de raíces. Las repeticiones serán eliminadas posteriormente.

La variable *steam* también guarda la raíz y será devuelta. Ahora bien, se iterará sobre todos los morfemas que contenga la palabra a analizar. La bandera *notgot* sera positiva si no se ha logrado encontrar la raíz, lo cual hará que se salte al análisis de postfijos. Si se ha llegado a más de la posición 17 de los prefijos, entonces el morfema, actual será la raíz, y se pasará a los postfijos. De lo contrario, se buscará el morfema en la posición actual. Si se encuentra, se agrega y se avanza a la posición siguiente. Si la no se encuentra un morfema, únicamente se avanza a la siguiente posición. Una vez alcanzada la raíz verbal, se utiliza el mismo procedimiento para encontrar postfijos.

La entrada de la función *segreg* será una palabra *verb*, en forma de cadena. En las variables iniciales se asignará al nivel *level* con cero, y a la bandera *preval* como verdadero, la cual indicará que nos encontramos recorriendo los prefijos, en caso de ser verdadera o los post fijos, en caso de ser falsa. También se tendrá una lista *seglist* donde se guardarán los morfemas encontrados y anotados y un archivo F que guardará todas las raíces.

Con los anteriores algoritmos podrán etiquetarse los corpus sementados, y se podrán extraer las palabras no algutinadas y las raíces. Estos dos últimos diccionarios son esenciales para un buen funcionamiento del segmentador y del traductor.

3.5. Interacción de los módulos

Ahora que se han explicado todas las partes componentes del sistema, se verá de manera más específica su implementación y el flujo de datos de las mismos. En el diagrama 3.9 es posible observar los módulos descritos: un corpus paralelo, una sección específica para preprocesar el idioma wixárika y el español, contenido en *wixarika.sh*, y una última parte que es el entrenador por frases *train-model.perl*. Ahora bien, en lo referente al proceso de traducción se utilizan los modelos generados, junto con *moses* para llevar acabo la traducción.

El corpus se divide en tres archivos diferentes. El primero es el recolectado por esta investigación donde una frase en wixárika descompuesta morfológicamente con antelación está alineada con una frase en español traducida. El segundo contiene un diccionario wixárika, que identifica sujetos, sus plurales y una traducción al español.

Por último se utiliza el corpus *Europarl*, para generar el modelo de lenguaje español.

Los tres corpus son insumos para el procesamiento. El corpus paralelo es separado y se generan dos archivos conteniendo las frases respectivas de cada idioma. A cada idioma se le trabaja según sus necesidades. En español únicamente se realiza un tokenizado y normalización, mientras que en wixárika se normaliza y, con ayuda del diccionario, se hace el segmentado morfológico. Con el corpus *Europal* únicamente se toma la parte en español, la cual se tokeniza, se normaliza y se entrena con *lmplz*.

Una vez preprocesadas las dos partes del corpus y el modelo de lenguaje se procede a entrenar el modelo. Esto consta de nueve pasos, donde se prepara el corpus, se llama *mgiza*, se alinean las palabras, se generan tablas de traducción léxica, se extraen frases, se evalúan las mismas, se aprende del reordenamiento de palabras y se genera el archivo del modelo, que especificará todo lo necesario a *moses* para su funcionamiento.

En la decodificación, se cuenta con la interfaz web que enviará el texto a traducir, dependiendo del idioma se realizarán pasos concretos. Si la dirección de traducción es del español al wixárika entonces primero se invoca al decodificador y posteriormente se realizará una reconstrucción del wixárika. En el orden inverso, primero se segmenta las palabras wixárika y posteriormente se invoca al decodificador.

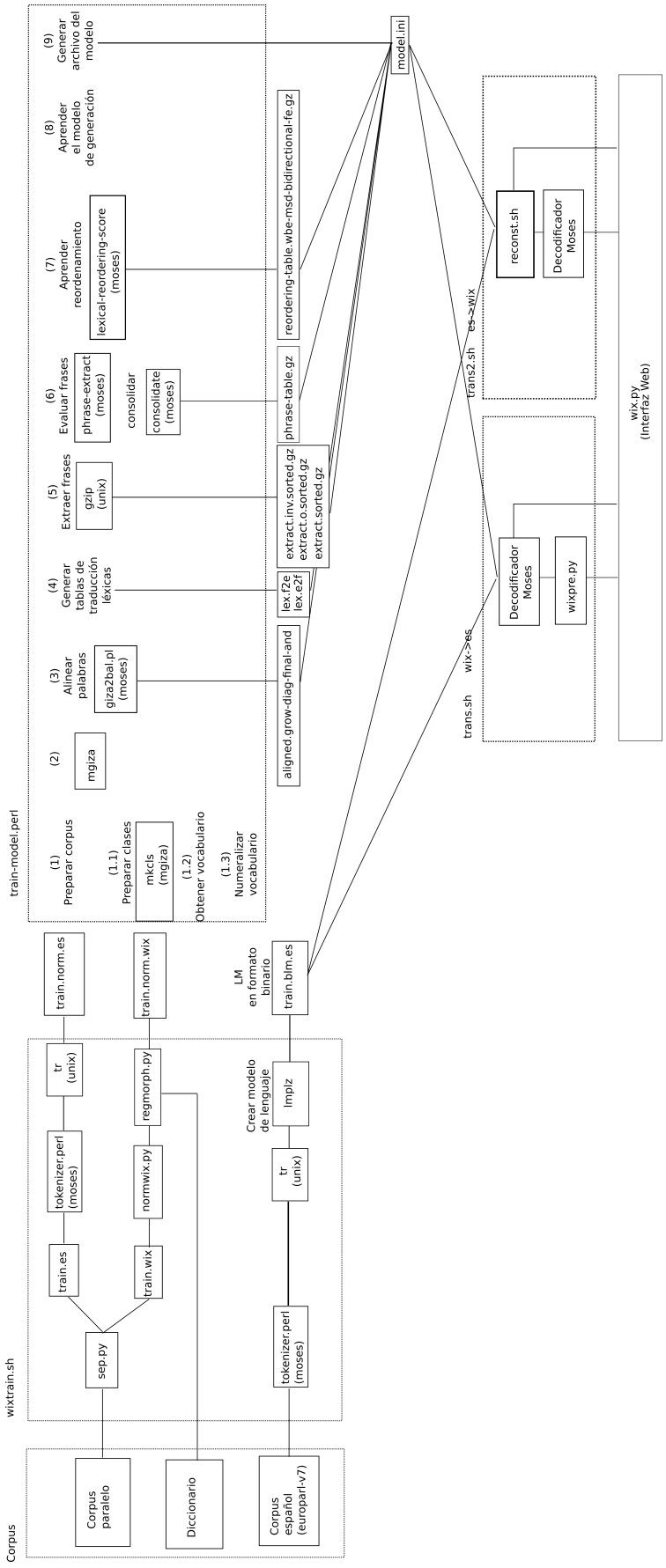

Figura 3.9: Diagrama de interacción de módulos

Para ampliar el corpus con correcciones o agregados, se ha implementado la capacidad de interacción con el usuario, que se explica en la figura 3.10. Los datos enviados por el usuario primero requieren que se analice si el texto es wixárika y español. En caso de ser positivo se envía a un archivo intermedio. Este tendrá que ser analizado por un humano hablante de las dos lenguas, para poder evaluar la calidad de la traducción. Las frases aptas serán ingresadas al corpus paralelo serán separadas y preparadas para un futuro entrenamiento. Una vez realizado esto, el administrador del sistema podrá reiniciar manualmente el entrenamiento, o se relega la tarea a un autoentrenamiento cada cierto tiempo.

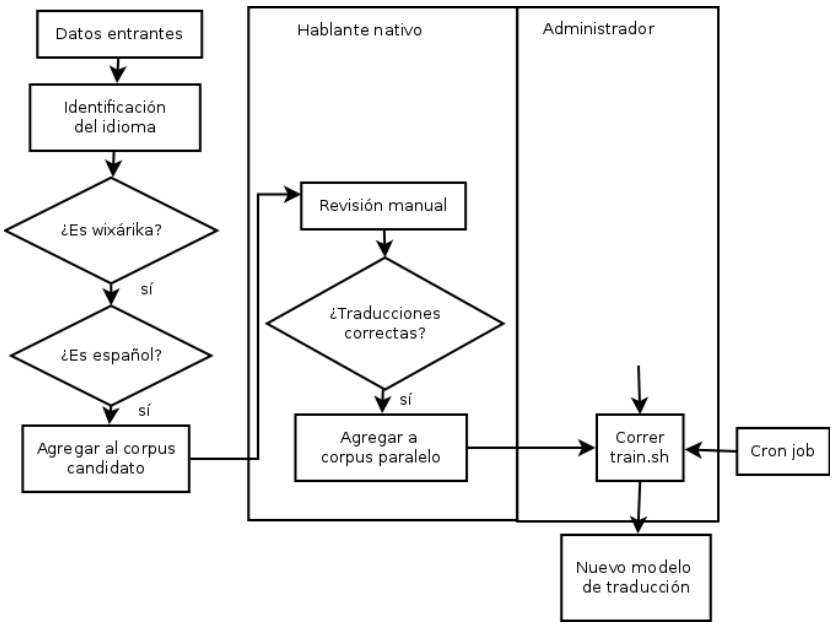

Figura 3.10: Diagrama de interacción de módulos

3.6. Interfaz web

La interfaz se construyó de tal forma para que sea sencillo realizar la traducción y corregir la misma. En la figura 3.12 se muestra la interfaz que se creó. Del lado izquierdo se contendrá el texto en wixárika, mientras que el texto español será contenido por la ventana derecha.

El usuario deberá introducir texto en el recuadro correspondiente al idioma origen. Posteriormente debe oprimir el botón de la dirección de traducción deseada y el sistema generará la traducción correspondiente en el cuando del idioma destino.

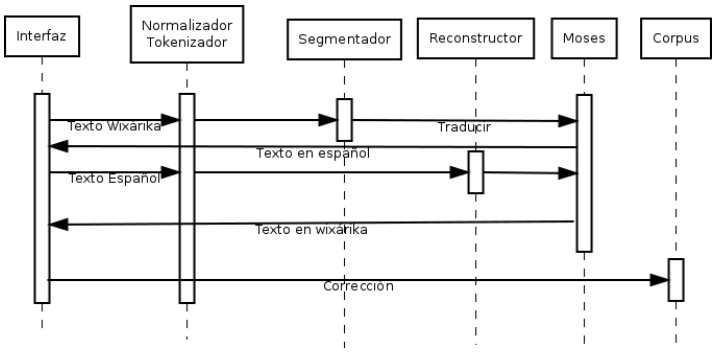

Figura 3.11: Diagrama de interacciones

Figura 3.12: Interfaz web

Para realizar una traducción, el usuario deberá corregir el error en los recuadros de los pares de idioma y presionar el botón de corrección. Con ello se enviarán los nuevos pares al servidor. Estos serán almacenados para su posterior evaluación e incorporación al corpus de entrenamiento.

También se proporciona un recuadro de ayuda con ejemplos de traducción, para personas no hablantes del wixárika. Claro está, siempre será posible traducir del español al wixárika.

Algoritmo 7 Analizador de palabras segmentadas

```
function SEGREG(verb)
    level ← 0
    preval ← 1
    List seglist ← ∅
    steam ← ∅
    File F ← OPEN(steamdic)
    para todo morph ∈ verb hacer
        nogot ← 1
        si preval = True entonces
            mientras nogot = True hacer
                si level > 17 entonces
                    e ← moprh
                    APPEND(seglist, e)
                    APPEND(steamdic, e)
                    level, preval, nogot ← 0
                    steam ← morph
                    Continue
                fin si
                si morph ∈ pre[level] entonces
                    e ← morph + level
                    APPEND(seglist, e)
                    level ← level + 1
                    nogot1
                si no
                    level ← level + 1
                fin si
            fin mientras
        si no
            mientras notgot = False  hacer
                si level > 22 entonces
                    devolver seglist
                fin si
                si morph ∈ post[level] entonces
                    e ← morph + str(−level)
                    APPEND(seglist, e)
                    level ← level + 1
                    nogot1
                si no
                    level ← level + 1
                fin si
            fin mientras
        fin si
    fin para
fin function
```

Capítulo 4

Resultados Obtenidos

Una vez descrita la metodología para el traductor se presentan los resultados obtenidos. A partir de un corpus paralelo se entrenaron tres traductores: el primero funciona con un modelo por frases SMT sin modificación alguna, el segundo con segmentación morfológica y el tercero con segmentación morfológica y etiquetado de los morfemas. Primero se explican las métricas de evaluación que se usaron en este texto, posteriormente se muestran los experimentos.

4.1. Evaluación

Para poder evaluar la calidad de la traducción se usan las métricas TER, WER y BLEU. Estas métricas sirven para comparar los resultados con otros trabajos. También en la plataforma de Internet se ha integrado una evaluación manual, con el objetivo de enriquecer el corpus y mejorar la traducción automática.

4.1.1. Manual

La evaluación manual es la forma más exacta de medir la calidad de una traducción. Como no existe una traducción óptima sino diferentes traducciones válidas, es difícil para un algoritmo evaluarlas. La única validación exacta es una evaluación humana, dado que un humano es capaz de distinguir entre un error de traducción y una traducción correcta. Sin embargo, estas evaluaciones requieren tiempo y personal que las realice.

En el sistema se llevará a cabo una evaluación humana por medio de la plataforma web, que permitirá al usuario contestar un cuestionario e incluso corregir la traducción.

La corrección será especialmente valiosa ya que permitirá ampliar el corpus del traductor y mejoraría los resultados en nuevos entrenamientos. Los datos ingresados deberán ser revisados para garantizar la validez de la nueva frase.

4.1.2. Automática

Para solucionar el problema del tiempo necesario en la evaluación humana se aspira a una evaluación automática que tenga alguna correlación con los humanos (Jurafsky & Martin 2000). En una evaluación no se puede tomar la magnitud del error como la distancia entre la traducción deseada y la obtenida, ya que un humano podría traducir de varias formas un mismo texto. Lo que se busca es encontrar una traducción que parezca humana. ¿Cómo lograrlo? Usando la mayor cantidad posible de traducciones humanas de una misma frase. La comparación entre la frase generada por la traducción y las frases humanas traducidas es la *distancia* entre los dos. Pero cómo medir esa distancia es el centro de la discusión.

WER

Word Error Rate(WER) es una métrica clásica que se toma prestado de los sistemas de reconocimiento de voz, para aplicarla a la traducción automática. Utiliza una distancia *Levenshtein* que es definida como el mínimo número de pasos de edición, contando inserciones, eliminaciones y sustituciones, que son necesarios para igualar la hipótesis con una traducción correcta (Koehn 2010). Tiene, por lo tanto, la desventaja de hacer referencia a la única traducción válida, pero por su sencillez permite plantear un primer acercamiento a la evaluación de una traducción.

$$WER = \frac{S + D + I}{N}, \tag{4.1}$$

donde S es el número de sustituciones, D es el número de eliminaciones, I las inserciones realizadas y N es el número de palabras en la hipótesis. Entre más alto sea el error, peor será clasificada la traducción. Para calcular el error, se utiliza programación dinámica.

La distancia *Levenshtein* se define como la distancia entre dos cadenas a, b, y es denotada por $\text{lev}_{a,b}(|a|, |b|)$,

$$\mathrm{lev}_{a,b}(a,b) = \begin{cases} \max(i,j) & \text{si } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{a_i \neq b_j}, \end{cases} & \text{de lo contrario} \end{cases}$$

donde $1_{a_i \neq b_j}$ es la función indicador que vale cero cuando $a_i = b_j$ es igual a 1, de lo contrario, $\mathrm{lev}_{a,b}$ es la distancia entre los primeros i caracteres de a y los primeros j caracteres de b.

BLEU

En BLEU se ordena cada traducción por la media de los pesos del número de n-gramas que concuerdan con la traducción humana. Una métrica de precisión simple de n-gramas tendría un sesgo al sobre valorar frases con palabras repetidas de alto peso. Para solucionarlo se usa una métrica de n-gramas de precisión modificada. Primero se computan los n-gramas que se emparejan, frase por frase. Después se agregan los contadores por cada frase candidata y es dividida entre el número de n-gramas en el corpus a probar, con el fin de computar la precisión modificada p_n, cómo se muestra a continuación:

$$p_n = \frac{\sum_{C \in \{Candidatos\}} \sum_{n-gram \in C} cont_{clip}(n-gram)}{\sum_{C' \in \{Candidatos\}} \sum_{n-gram' \in C'} cont(n-gram)}. \tag{4.2}$$

El método tiene, a su vez, problemas con frases cortas, por lo que es necesario realizar una penalización. Sea c el tamaño de la frase y r el tamaño efectivo del corpus de referencia, computamos la penalización llamada BP (Papineni et al. 2002).

$$\alpha(x,y) = \begin{cases} 1 & \text{si } c > r \\ e^{(1-\frac{r}{c})} & \text{si } c \leq r. \end{cases} \tag{4.3}$$

Con el valor obtenido se puede obtener la métrica BLEU que nos permitirá cuantificar la evaluación. En la experiencia se ha descubierto que $n = 4$ es más preciso con pesos

uniformes $w = 1/N$ (Papineni et al. 2002).

$$\text{BLEU} = \text{BP} \exp(\sum_{n=1}^{N} w_n \log p_n) \tag{4.4}$$

$$\log \text{BLEU} = \text{mín}(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n. \tag{4.5}$$

Se lleva la ecuación 4.4 al espacio logarítmico para poder visualizar mejor el valor generado, cómo se muestra en 4.5.

TER

Translation Edit Rate(TER) (Snover et al. 2006*a*) es una métrica automática que mide la cantidad de edición requerida por un humano para cambiar la salida generada por el sistema con el objetivo de llegar a ser igual a la traducción de referencia. Cada edición tiene un costo y las operaciones existentes son insertar, eliminar, intercambio y sustitución. La métrica, por lo tanto, es el costo de edición mínimo encontrado de una frase traducida por un sistema en comparación con una frase referencia, traducida por un humano.

$$TER = \frac{\text{número de ediciones}}{\text{número promedio de palabras en la referencia}} \tag{4.6}$$

El algoritmo tiene dos desventajas. El hecho de que una frase origen puede tener varias frases destino, y por lo tanto TER valuaría de manera errónea la traducción; y el problema de que se ha comprobado que calcular la distancia de edición con operadores de movimiento es un problema *NP*-Completo(Snover et al. 2006*a*) obligando a utilizar algoritmos de aproximación, que se puede ver en el algoritmo 8.

Con el algoritmo 8 se logra encontrar el número de ediciones, el cual será sustituido en la ecuación 4.6. Entre menos sea el valor de TER mejor será el resultado. Así que se buscará minimizar este valor. El aspecto positivo de este método es que no requiere de frases largas para poder ser evaluado, lo cual mejora su comportamiento en traducciones que miden el wixárika. Al aglutinar la información de traducción, el wixárika podría traducir una frase de diez o más palabras en español en una única palabra, lo cual no podría ser evaluado en BLEU.

Algoritmo 8 Calcular el número de ediciones

Entrada: Hipótesis h, Referencia R
$E \leftarrow \infty$
para todo $r \in R$ **hacer**
  $h' \leftarrow h$
  $e \leftarrow 0$
  repetir
    Encontrar un cambio s que reduzca $\text{mín}_{edist}\, h', r$
    si s reduce la distancia de edición **entonces**
      $h' \leftarrow$ aplicar s a h
      $e \leftarrow e + 1$
    fin si
  hasta que No quedan intercambios que reduzcan la distancia $e \leftarrow e + \text{mín}_{edist}(h', r)$
fin para

4.2. Experimentos

Las pruebas se realizaron en una computadora con dos procesadores Intel Xeon X3450 x86 de 64 bits con 4 núcleos cada uno y capacidad de dos hilos por núcleo, a 2.67 GHz NUMA, con 16 GB de memoria RAM. Para el sistema de alineado se usó *GIZA++* (Och 1999) y para la extracción de la tabla de frase y decodificación se usó el sistema Moses (Koehn et al. 2003). El corpus fue extraído del libro (Gómez 1999) que aporta valiosa información morfológica en su texto apareado. En la tabla 4.1 se muestra las características del corpus alineado con el que se entrenó el traductor.

El traductor en línea y la interfaz web se ejecuta en el mismo servidor antes descrito y utiliza un servidor web Apache/2.4.10 con WSGI, con el *framework* Flask 0.10.1 y Python 2.7.9. Para alinear palabras, se utiliza MGIZA; para el entrenamiento de frases y decodificación Moses compilador desde git (con hash mmt-mvp-v0.12.1-743-gea306f6). Para la compilación se usaron las bibliotecas Boost versión 1.59.0 y Xmlrpc-c versión 1.33.17. Para el enlace entre las herramientas creadas, Moses y MGIZA, se implementaron una serie scripts Bash(GNU bash, version 4.3.30).

	español	wixárika
Lineas	790	790
Palabras	3810	2347
Tokens	874	1197
Tamaño	20 KB	22 KB

Tabla 4.1: Corpus usado

El tamaño del corpus es muy reducido, si se compara con el corpus Europarl (Koehn 2005) que contiene en sus idiomas más estudiados aproximadamente 2 millones de frases con alrededor de 50 millones de palabras en inglés y 44 millones de palabras en el idioma origen. Los idiomas con menor corpus contienen de 300 mil a 700 mil frases con 10 millones de palabras en inglés e igual número de palabras en el idioma origen. Para mejorar el rendimiento del corpus utilizado para el presente trabajo se usó segmentación morfológica, pero se espera una fuerte penalización en rendimiento comparado con los sistemas que son entrenados con grandes cantidades de datos.

4.3. Prueba de concepto

Para la prueba de concepto, únicamente se utilizaron 100 frases apareadas como corpus de experimentación para el entrenamiento del sistema, y las traducciones se realizaron con los morfemas y las palabras usadas en el mismo corpus. También se contó con un texto segmentado de entrada, así como para el entrenamiento, por lo que no se tomó en cuenta el error de segmentación. Las frases con las cuales se evaluó el traductor fueron simples y de tamaño reducido. El objetivo fue comprobar, si la segmentación influía positivamente en la traducción, o esta propuesta disminuía su rendimiento.

Para la evaluación no se midió con BLEU (Papineni et al. 2002) por el tamaño de las frases usadas, y se prefirió WER (Zechner & Waibel 2000) y TER (Snover et al. 2006*b*), que son eficientes en estas condiciones. Con los valores obtenidos se realiza una comparación de resultados (ver tabla 4.2), entre una traducción sin segmentación, con segmentación y con etiquetado.

	WER	TER
Sin segmentación morfológica(SGM)	38	0.84
Con segmentación morfológica(CSM)	25	0.46
Segmentación con etiquetado(CSEM)	21	0.46

Tabla 4.2: Evaluación de traducción

El error en la traducción automática usando palabras sin segmentación es más alto que si usamos un segmentador morfológico. El problema que se encuentra en una traducción normal es que se necesita forzosamente encontrar la misma combinación de morfemas aglutinados en torno a una raíz verbal de la frase a traducir en el texto de entrenamiento; mientras que en el modelo con segmentación, se entrena al traductor la forma en que se deben realizar las combinaciones de morfemas wixaritari para que generen ciertas frases en español. Los resultados, usando además un etiquetador de morfemas, son ligeramente superiores al hecho de no usarlo (en el caso del español al wixárika). Este resultado se debe a la desambiguación de morfemas que se encuentran en posiciones diferentes dentro del verbo o sustantivo, con la penalización de agregar mayor error cuando no es posible traducir la palabra.

wixárika	Sin segmentar	Segmentado
neki	neki	mi casa
'aki p+tuxa	'aki es blanca	tu casa blanca
hakewa ne ki	esta falta es no es nueva	esta falda no es nueva

Tabla 4.3: Ejemplos de traducción

En la tabla 4.3 se muestra una comparación entre traducciones simples, mostrando las deficiencias de los modelos usados. La palabra *'aki*, al no encontrase explícitamente en el corpus de entrenamiento, no pudo ser traducida por el modelo SMT sin segmentación. Pero esto es solucionado al entrenar con segmentación, encontrando la descomposición *'a* y *ki*, e identificando el primero con la palabra *tu* y el segundo con *casa*. El reto, en la traducción con segmentación, es entrenar al traductor de tal forma para que aprenda el funcionamiento de la aglutinación del wixárika. En contraste, la traducción sin segmentación buscará encontrar correspondencias por palabras.

Este primer experimento presenta indicios de la viabilidad de la traducción con la metodología propuesta, y encuentra una mejora significativa usando la segmentación.

Con estos resultados, se pasó a ampliar el corpus y a implementar un segmentador para el uso en texto sin tratamiento previo.

4.4. Wixárika a español

Para la traducción del español al wixárika se utilizó un corpus 790 frases. De estas 790 se extrajeron 50 frases de manera aleatoria para usarlas como traducción prueba, y se realizaron los experimentos presentados a continuación. Para evaluar fueron aplicadas las métricas WER, TER y BLEU, con el fin de compararlas con resultados de otras traducciones automáticas.

En la tabla 4.10 se muestra nuevamente una comparación entre las tres distintas metodologías con las cuales se experimentó en la prueba de concepto.

	WER	TER	Bleu
Sin segmentación morfológica(SGM)	72	0.8875	6.38
Con segmentación morfológica(CSM)	58	0.6625	25.19
Segmentación con etiquetado(CSEM)	58	0.6625	23.69

Tabla 4.4: Evaluación de traducción wixárika a español

El error se incrementó de manera importante con respecto al primer experimento debido a que al introducir mayor número de frases se presentaron nuevas y más complejas formas morfológicas y sintácticas. Las frases con las cuales se entrena ya no son tan sencillas como las iniciales. Sin embargo, se sigue observando la tendencia, en la cual la traducción mejora usando segmentación morfológica, y empeora al prescindir de ella. A continuación se muestran ejemplos de traducción con errores identificados, de cada metodología de traducción.

Sin Segmentar	Segmentado	Traducción humana
no es blanca neki	en mi casa no es blanco	mi casa no es blanca
el pájaro p+kawaiya	no es el pájaro gordo	ese pájaro no está gordo
pep+kakutsu	nosotros somos es de ellos gordo	ellos son gordos

Tabla 4.5: Ejemplos de traducción y sus dificultades

Como se puede observar en la tabla 4.5, la traducción que no utiliza segmentación tiene gran dificultad de encontrar todas las combinaciones posibles de aglutinación en

las palabras wixaritari. Conforme se incrementa la complejidad de la frase a traducir, las combinaciones se incrementan y se tenderá a encontrar menos palabras en los datos de entrenamiento. Por el otro lado, la segmentación morfológica en el último ejemplo tiene fuertes dificultades para aprender de manera correcta las complejas reglas del wixárika. Si bien se logra acercar en cierta medida a la traducción deseada, es necesario un corpus más grande para poder entrenar al sistema con más estructuras morfológicas. También se encontró un problema en el orden de las palabras generadas en español, problema que aparece en las dos primeras frases. Esto es característico de un pobre LM. Si bien el corpus usado (Europal) es amplio, este no incluye un lenguaje cotidiano y se reduce a las transcripciones de los debates parlamentarios. Pero el lenguaje cotidiano es la temática del corpus wixárika. Por lo anterior, la corrección mediante el LM casi no se refleja en una mejor calidad de traducción en estos ejemplos.

Para tener un referente con otros experimentos, se retoma el trabajo de Koehn (Koehn 2005) que utiliza el corpus Europal, donde se presentan valores BLEU para el corpus de grandes dimensiones, utilizando traductores por frases. El mejor par de lenguas traducidas es español al francés con 40.3 y francés al español con 38.4 BLEU, seguido del par portugués al francés con 39 BLEU y en la dirección contraria 35,9. Los experimentos, donde se traduce inglés como fuente, tiene resultados inferiores, con sus mejores desempeños a español 30,1 y francés 31,1 y resultados semejantes en la dirección inversa. Koehn llega en este trabajo a la conclusión de que los lenguajes más cercanos entre ellos tienen mejores resultados, mientras que los idiomas más distantes tienen mayores dificultades. Los peores resultados obtenidos fueron entre todas las lenguas con respecto al finlandés, con BLEU bajos hasta de 10.3 (holandés a finlandés) y casos semejantes a otros idiomas. El caso del finlandés (de la familia lingüística urálica) es importante para el wixárika, ya que comparten morfologías complejas y aglutinantes.

Utilizando los ejemplos del libro Hablemos Español y Huichol (José 2009), texto que recopila frases sencillas hechas para la enseñanza de los dos idiomas, se han realizado traducciones para mostrar su comportamiento con algunas estructuras morfológicas del wixárika con sus equivalentes sintácticas en español.

Wixárika	Español
ne nep+'uki	yo soy un hombre
ne nep+'uka	yo soy mujer
ne nep+temaik+	yo soy un muchacho
ek+ pep+'uka	usted es la mujer
ek+ pep+'uki	yo soy un hombre
tame tep+'uki	nosotros somos el hombre
xeme xep+uka	ustedes son mujer

Tabla 4.6: Ejemplos: yo soy (wixárika a español)

La forma en que se construyen las palabras wixárika para referirse *yo soy*, es la siguiente: en primer lugar la palabra *ne* (*ne-*) es el pronombre personal yo. Los restantes pronombres son *'ek+* (*pe-*) para la segunda persona singular, *m+k+* para la tercera persona plutar, *tame* (*te-*) primera persona plural, *xeme* (*xe-*) segunda persona plural y *m+me* (*me-*) o sin usar para la tercera persona plural. La segunda palabra es el verbo, en el primer ejemplo se descompone en los morfemas *ne-p+-'uki*. *'uki* es la raíz verbal, que significa hombre. El morfema *ne-* retoma la primera persona y *p+* es el asintor general del lenguaje. En la tabla 4.11 se muestran ejemplos de traducción con las raíces *'uki*, *'uka* (mujer) y *temaik+* (muchacho).

Para calificar al sujeto, es posible incorporando esta característica al verbo. El sujeto se escribe anterior al verbo. En este caso se utiliza *'uki* y *huku* (pino). La palabra *'ik+* significa *este*. Para declarar si se es alto o pequeño se utilizan la raíz *tewi*. Para calificar como alto al sujeto se utiliza la descomposición *'a-p-u-tewi*, donde *'a-* es el posesivo de segunda persona, *p-* es el asintor *p+-* y *u-* refiere visibilidad. Para denotar que el sujeto es pequeño se utiliza el diminutivo *'e-* y *tsi-*. En el ejemplo de traducción (tabla 4.12)

Wixárika	Español
'ik+ 'uki 'aputewi	este hombre es alto
'ik+ 'uki 'etsitewi	este hombre es la
'ik+ huku 'aputewi	este pino es alto
'ik+ huku 'etsiputewi	este pino es chaparro

Tabla 4.7: Ejemplos: Calificativos de altura (wixárika a español)

De la misma forma se traducen correctamente algunas partes del cuerpo descritas por

Grimes. Las referencias a la cabeza y la mano, combinados con pertenencia en singular se muestran en la tabla 4.8. Como se ha mostrado anteriormente se descompone el verbo y se traduce por morfemas, para poder aprender las funciones da cada uno. Las raíces en este caso son *mu'u* para cabeza y *mana* para mano.

Wixárika	Español
'ik+ mep+mu'u	esta es mi cabeza
m+k+ 'ap+mu'u	es es tu cabeza
'ik+ nep+mana	esta es mi mano
'ik+ 'ap+mana	este es tu mano
'ik+ p+mu'uya	esta es su cabeza

Tabla 4.8: Ejemplos: Pertenencia de partes del cuerpo (wixárika a español)

Una información importante para la traducción, que el wixárika no contiene, es el género. Esto dificulta la traducción al español. En el ejemplo de la tabla 4.14 se logra ver la dificultad de la traducción, misma que incluso los hablantes del wixárika presentan al no practicar de manera habitual el español. La palabra *m+k+* se usa como pronombre en tercera persona singular sin considerar el género. Lo mismo sucede con *m+me*, tercera persona plural. En ambos casos, es imposible para el traductor intuir el género de la o las personas. Por otro lado, la localización de la persona, en este ejemplo, se realiza por medio de las palabras *'ena* (aquí) y *'uma* ahí. La palabra *puwe* significa *estar parado*. Al momento de aglutinarse, se convierte en la raíz *'u*.

Wixárika	Español
m+k+ 'ena puwe	ella está parado aquí
m+k+ 'uma puwe	ella está parado ahí
m+me 'ena mep+ti'u	ellos están parados aquí
m+me 'uma mep+ti'u	ellos están parados ahí
xeme 'uma xep+ti'u	ustedes están parados ahí
m+k+ 'uma puwe	ella está parado ahí
tame 'ena tep+ti'u	nosotros estamos aquí parados

Tabla 4.9: Ejemplos de traducción y sus dificultades (wixárika a español)

Con estos sencillos ejemplos de traducción es posible ver la capacidad del sistema,

para realizar su tarea sobre frases simples en apoyo al aprendizaje del idioma. Una gran ventaja que ofrece, es la descomposición automática de las palabras aglutinadas.

4.5. Español a wixárika

La dirección inversa, del español a wixárika, presenta mayores dificultades. El mismo problema también se encuentra en los experimentos de Koehn (Koehn 2005) con el finlandés. La traducción de una lengua fusionante a una lengua aglutinante es de gran dificultad. En la tabla 4.10 se muestra que el BLEU sin segmentación baja hasta 5.77 y mejora con segmentación y etiquetado hasta 7.37. La segmentación con etiquetas además utiliza una reconstrucción de la palabra wixárika. Sin embargo, con segmentación simple, se presenta un nuevo problema: ¿qué criterio se tiene que tomar para distinguir raíces de morfemas y palabras no aglutinadas, y cómo se debe determinar la forma de aglutinar morfemas? El etiquetado permite identificar los morfemas y establecer hacia que dirección deben ser aglutinados. La palabra no etiquetada que se encuentre en medio de un grupo de morfemas será asumida como raíz. Es por esta razón que la calidad de traducción morfológica sin etiquetas presenta errores grandes.

	WER	TER	Bleu
Sin segmentación morfológica(SGM)	49	1.088	5.77
Con segmentación morfológica(CSM)	86	1.91	0
Segmentación con etiquetado(CSEM)	39	.866	7.37

Tabla 4.10: Evaluación de traducción español a wixárika

El caso de estudio no es para minimizar. El finlandés, con una morfología menos compleja que el wixárika y con corpus más grandes, tiene un BLEU que varían entre 10.3 (holandés a finlandés) y 15 (sueco a finlandés). El problema es atribuido por Koehn a que el finlandés tiene una alta complejidad morfológica y es bastante aglutinante. Se considera que en general, es más sencillo traducir de una lengua rica en información a una que contiene poca información y es más difícil en la dirección contraria. Lo mismo ha sido observado en la traducción árabe-inglés.

A continuación se presentarán los mismos ejemplos del libro de Grimes (José 2009), ahora en la dirección de traducción español a wixárika. Se comienza con el ejemplo sobre pronombres personales en la tabla 4.11. Estos ejemplos sencillos se traducen al wixárika fácilmente.

Wixárika	Español
ne nep+'uki	yo soy hombre
ne nep+'uka	yo soy mujer
ne nep+temaik+	yo soy un muchacho
ne nep+'+rimari	yo soy una muchacha
ek+ pep+'uka	tu eres mujer
ek+ pep+'uka	usted es mujer
ek+ pep+'uki	tu eres hombre
ek+ pep+'uki	usted es hombre
tame tep+'uki	nosotros somos hombres
xeme xep+'uka	ustedes son mujeres

Tabla 4.11: Ejemplos: yo soy (wixárika a español)

En el siguiente ejemplo, que describe características de un sujeto, se aprecia una confusión del traductor entre dos raíces verbales muy parecidas. La raíz *tewi* describe una característica de altura, mientras que *pawi* de tamaño en general. Para alto, las frases son idénticas, pero para pequeño, cambia el sentido de la frase meta.

Wixárika	Español
'ik+ 'uki 'aputewi	este hombre es alto
'ik+ 'uki 'etsipawi	este hombre es chaparro
'ik+ huku 'aputewi	este pino es alto
'ik+ huku 'etsipawi	este pino es chaparro

Tabla 4.12: Ejemplos: Calificativos de altura (wixárika a español)

A continuación se muestra en la tabla 4.13 ejemplos de traducción sobre partes del cuerpo y pertenencia. La traducción se da correctamente, usando la aglutinación de los morfemas necesarios.

Wixárika	Español
'ik+ p+mu'u	esta es mi cabeza
'ik+ 'ap+mu'u	esta es tu cabeza
'ik+ nep+mana	esta es mi mano
'ik+ 'ap+mana	este es tu mano
'ik+ p+mu'uya	esta es su cabeza

Tabla 4.13: Ejemplos: Pertenencia de partes del cuerpo (wixárika a español)

También en la traducción referente a localización, el traductor ha logrado generar de forma correcta los morfemas y ordenarlos. El problema de mala traducción de géneros no existe en esta dirección, ya que no son tomados en cuenta por el wixárika.

Wixárika	Español
m+k+ 'ena puwe	el está parado aquí
m+k+ 'uma puwe	el está parado ahí
m+me 'ena mep+ti'u	ellos están parados aquí
m+me 'uma mep+ti'u	ellos están parados ahí
xeme 'uma xep+ti'u	ustedes están parados ahí
m+k+ 'uma puwe	el está parado ahí
tame 'ena tep+ti'u	nosotros estamos parados parados

Tabla 4.14: Ejemplo: localización (wixárika a español)

4.6. Comparación de resultados

El primer problema, que se enfrenta la traducción wixárika-español, es el bajo corpus. Koehn (Koehn et al. 2003) demuestra que los resultados mejoran conforme se incrementa el corpus paralelo en SMT. En un experimento de alemán a inglés los resultados mejoran constantemente conforme se incrementa el corpus, a partir 10 mil frases. En el caso del wixárika-español, la cantidad de corpus escaso representa un primer reto.

El problema, además, se incrementa por la gran distancia entre idiomas y, en particular, por la característica aglutinante del wixárika. En la tabla 4.15 se presenta una comparación de diversos resultados obtenidos en trabajos previos sobre características relevantes para el caso de estudio.

En primer lugar se presenta una traducción considerada de bajos recursos entre dos de las lenguas más estudiadas, basada en el corpus Europal. La traducción alemán a inglés obtuvo un BLEU de 22.5. Los mejores resultados para estos idiomas se consiguen con corpus de millones de frases. Para el mismo par de idiomas, con un corpus grande, se obtuvo un BLEU de 29.3 (Koehn et al. 2003). El caso del finlandés, con un corpus mediano, se muestra un bajo rendimiento. Para el par inglés a finlandés se obtuvo un bajo rendimiento, dada a la problemática de traducir a una lengua aglutinante. Los resultados para este caso son semejantes en el caso del turco. Sin embargo, Oflazer (Oflazer 2008) plantea un metodología especial para tratar el problema, mejorando BLEU hasta 24.69, con 45 mil frases.

Número de frases	Bleu	dirección	Comentario
1,023,523	30	francés a inglés	SMT Estado del Arte (Koehn 2005)
10 000	22.5	alemán a inglés	SMT (Koehn 2005)
941,890	21.8	finlandés a inglés	SMT de un lenguaje aglutinante a uno fusionante (Koehn 2005)
941,890	13	inglés a finlandés	SMT de un lenguaje fusionante a uno aglutinante (Koehn 2005)
45,709	16.13	inglés a turco	SMT de un lenguaje fusionante a uno aglutinante (Oflazer 2008)
45,709	24.61	inglés a turco	SMT de un lenguaje fusionante a uno aglutinante con trabajo morfológico (Oflazer 2008) además de correcciones al español.
790	25.19	wixárika a español	SMT de un lenguaje polisintético a uno fusionante con segmentación morfológica
790	7.37	español a wixárika	SMT de un lenguaje polisintético a uno fusionante con segmentación morfológica y etiquetado

Tabla 4.15: Comparación con otros trabajos

El caso de estudio del presente trabajo, al conjuntarse un idioma de aglutinante, en el caso del wixárika, escasos pares de frases alineadas, se logra obtener un BLEU de

25.19 del wixárika a español y 7.37 del español a wixárika. En el experimento habrá que considerar posibles sesgos por el reducido corpus de experimentación. En los resultados del presente proyecto, habrá que agregar otras dificultades no consideradas, como diferentes escrituras: la concepción no uniforme de que es una palabra por parte de los hablantes, y la necesidad de recolectar y experimentar con un corpus más grande.

4.7. Guía de uso

La interfaz de usuario presenta un traductor sencillo entre los dos idiomas, con dos campos, de introducción de texto. En el primer campo se espera la entrada de un texto wixárika, lo cual se constata por medio del identificador del lenguaje. En el segundo campo de entrada se espera un texto en español. Entre los dos campos de texto, se encuentran dos botones con el sentido de la traducción. En la parte inferior están los botones auxiliares.

Figura 4.1: La interfaz gráfica

Para traducir wixárika a español se debe introducir el texto origen wixárika en su campo de texto, y posteriormente presiona el botón del sentido hacia la derecha, cómo se muestra en la figura 4.2.

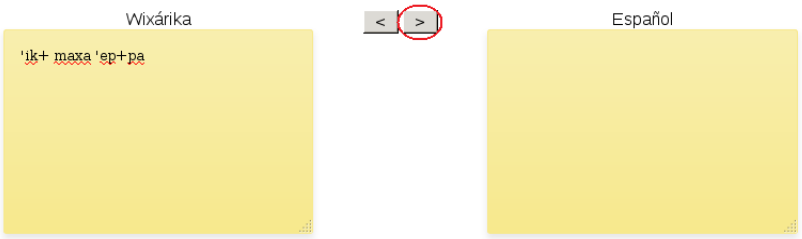

Figura 4.2: Traducción del wixárika al español

El resultado de la traducción aparecerá en el campo de texto dedicado al español. Es posible que exista un retraso de unos segundos, necesarios para realizar el proceso de segmentación y traducción.

Figura 4.3: Traducción wixárika a español exitosa

En el sentido español a wixárika se debe introducir el texto en español en el recuadro derecho y presionar el botón de traducción izquierdo, como se muestra en la figura 4.4

Figura 4.4: Traducción español a wixárika

Una vez concluido el proceso de traducción aparecerá en el recuadro wixárika la traducción, como se muestra en la figura 4.5

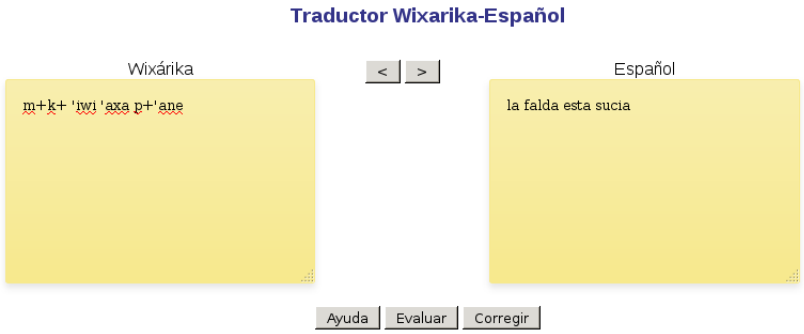

Figura 4.5: Traducción español a wixárika exitosa

Para corregir la traducción se requiere corregir el texto en los recuadros amarillos. Para enviar el nuevo texto se puede presionar el botón de corregir, el cuál será enviado al servidor para ser almacenado y posteriormente ser supervisado por un humano con el fin de ser incorporado al corpus principal.

Por último, para conveniencia del usuario, se proporciona una ventana de ayuda para el usuario que quiere utilizar la plataforma. También se incluyen ejemplos en wixárika para probar el traductor.

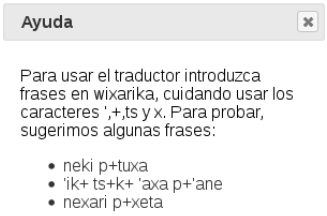

Figura 4.6: Ventana de ayuda

Capítulo 5

Conclusiones y trabajo futuro

En el presente trabajo se ha implementado un traductor del wixárika al español y del español al wixárika, basado en traducción estadística automática (SMT). El traductor se enfrenta al reto de contar únicamente con una cantidad limitada de ejemplos de traducción, es decir para esta combinación existen pocos recursos, tanto de corpus como de conocimientos gramaticales. El corpus escaso impide utilizar los sistemas SMT sin realizar modificaciones a su proceso de entrenamiento y traducción. También el escaso estudio de gramática y morfología del wixárika impide que se tome el modelo de traducción por reglas. Utilizar el modelo RBMT limitaría automáticamente el traductor a un único par de idiomas, sin ampliar el caso a otros pares de idiomas indígenas. La gran ventaja de utilizar SMT para la traducción es que para idiomas semejantes no se requiere mayores modificaciones a la metodología, sino únicamente un corpus alineado del par de idiomas a traducir. El texto puede ser recopilado de libros u otras fuentes o creado específicamente para el traductor.

El traductor, que se presenta, es el primer trabajo de traducción automática y NLP para el wixárika. Trabajos semejantes han sido realizados para otras lenguas indígenas. *Microsoft Translator Community Partners* (Microsoft 2016) es un proyecto de código y datos cerrados para el ñañú (otomí) de Querétaro y el maya de Yucatán. El hecho de ser un proyecto cerrado impide la amplia experimentación y un aporte general de la comunidad. El proyecto Apertum (Forcada et al. 2011) fue extendido para traducir quechua al español y viceversa por Calderón (Calderón et al. 2009). El traductor mencionado utiliza RBMT, lo cual lo restringe al idioma trabajado. Para los idiomas yutonahuas, un trabajo relevante ha sido el realizado por Gutiérrez (Gutierrez-Vasques et al. 2016) al recopilar un corpus paralelo en náhuatl-español, tanto de náhuatl clásico, como de

moderno, sumando 1,186,662 tokens entre los dos idiomas. Con el sistema SMT de este trabajo y el corpus recolectado para otros idiomas, se abre la puerta para que nuevos pares de idiomas puedan ser traducidos, sobre todo si son semejantes como el wixárika y el náhuatl.

En el capítulo de resultados se muestra la capacidad de traducción que tienen los SMT del estado del arte y se compara el corpus existente y sus resultados para estos traductores. Durante la experimentación, la implementación y la comparación con otros trabajos, fue posible ver cuatro grandes retos y un corolario para la traducción SMT de un idioma como el wixárika:

- **Problema de la traducción con recursos escasos**. Cómo ya se ha planteado al inicio de este trabajo, los idiomas indígenas u originarios, carecen de amplias fuentes escritas que permitan su análisis. Los pocos recursos de los que se dispone deberán, por lo tanto, ser aprovechados de tal manera que se obtengan resultados aceptables.

- **Poca estandarización del lenguaje y su escritura**. Si se toma el texto en wixárika, u otros idiomas indígenas, el análisis del lenguaje se enfrenta a una gran cantidad de ruido proveniente de diferentes escrituras, ortografías, conceptos de palabras, y sobre todo de dialectos dentro del mismo idioma. Si bien un normalizador y un tokenización reducen este ruido, no logran complementar la falta de información en ciertas escrituras o las diferencias entre dialectos del mismo lenguaje. En el trabajo únicamente se trató la variante del wixárika de San Andrés Comiatha (*tateikie*), con una tabla de equivalencia de signos creados con base a la observación del uso moderno en redes sociales de los hablantes y comparándolos con diversos textos del idioma.

- **La traducción entre lenguas distantes**. Un problema en la traducción automática es la traducción entre idiomas gramaticalmente lejanos. De manera regular, los lenguajes con mayor similitud tienen mejores resultados (Koehn 2005), formando grupos. El caso del wixárika con el español cae en el caso de lenguas distantes. Los dos idiomas son muy diferentes. El español es una lengua fusionante, de la familia indo-europea, y el wixárika, una lengua polisintética y aglutinante, de la familia yutonahua. Esta distancia dificulta la traducción.

- **Traducción de lenguas algutinantes**. Lenguas como el wixárika, la familia yutonahua en general y gran parte de las lenguas indígenas del continente americano

son aglutinantes. Esto implica que en torno a una raíz, se aglutinan morfemas que agregan significados a esta raíz. La información aglutinada de esta forma puede llegar a ser muy amplia. Las lenguas fusionantes, por el contrario, no tienen esta capacidad y expresan menos información en sus frases. La traducción entre estas lenguas es complicada por tener topologías distintas y por contener cantidades diferentes de información por palabra, lo cual complica la alineación. Si bien, en este trabajo se ha demostrado que una alineación palabra-morfema ayuda al proceso de alineamiento y traducción, no existe un equivalente entre ellos.

- **El caso especial de traducción de una lengua con poca riqueza morfológica a una algutinante**. En el caso de la traducción de un lenguaje aglutinante a un lenguaje fusionante, existe una importante pérdida de información. Este problema se ha registrado en varios estudios (Koehn 2005, 2010, Oflazer 2008). En el caso del wixárika a español, se encontró el mismo problema, con una traducción inferior del español a wixárika, que el del wixárika a español.

Dado las dificultades expuestas, se presentó una metodología de un traductor estadístico haciendo uso de información morfológica previa, que permite usar de mejor manera el reducido corpus existente. Dado que una lengua aglutinante forma sus palabras mediante reglas morfológicas y un conjunto de morfemas, se entrena al traductor con estas reglas. Un traductor SMT sin modificaciones intenta traducir con palabras aglutinadas. Para llevar a cabo esto de manera correcta, se necesitarían frases en el corpus que contengan todas las posibles combinaciones de morfemas. Sin embargo, dos aspectos no hacen práctico esta posibilidad: la falta de un corpus tan grande, y que los hablantes forman libremente palabras con las reglas morfológicas, por lo que un corpus no puede abarcar todas las combinaciones. En las experimentaciones se logró observar este fenómeno con transferencia de importantes cantidades de palabras sin traducir. Para la traducción del español a wixárika se tuvo que implementar un aglutinador que une los morfemas generados con un traductor entrenado mediante morfemas etiquetados. Comparando los resultados, tomando en cuenta las limitantes de bajos recursos, los resultados obtenidos son el primer paso para un amplio estudio de estos lenguajes en el marco del NLP. Los valores BLEU, WER y TER, incluso superan a experimentos realizados con lenguas con mayor corpus, pero evidentemente, no llegan a tener la calidad de los pares de idiomas que cuentan con millones de frases apareadas y que son cercanos entre ellos.

Las tareas de segmentado y etiquetado son esenciales para el modelo. La implementación de un FST para este fin logró buenos resultados. Esta herramienta es de gran importancia, dado que una mala segmentación lleva a un mal entrenamiento y una mala traducción. Decidir entre diferentes opciones de segmentación también es un reto importante a resolver.

Para el presente proyecto se crearon las siguientes herramientas: normalizador, tokenizador, segmentador morfológico, identificador y extractor del wixárika, un sistema de entrenamiento y prueba para Moses, y una plataforma web[1]. A esta página se incorporó la traducción en los dos sentidos, un corrector y un evaluador manual. La información que se obtenga de la plataforma servirá para mejorar la traducción y ampliar el corpus.

El desarrollo con el paradigma de software libre provee una herramienta wixárika-español para los fines que las personas de los pueblos y comunidades requieran, disponible para su uso en una plataforma web y con los recursos de NLP, abiertos para su uso y modificación bajo licencia GPL[2]. Liberar estas herramientas, permite aportar mejoras, tanto al corpus como a las herramientas.

Para enriquecer y mejorar el proyecto, se han realizado una serie de presentaciones en congresos y eventos, además de la publicación de dos artículos.

- Plática "Traductor Wixárika-Español" en el marco del *Festival Galois, primavera 2016*, el 24 de mayo de 2016, en la UAM Azcapozalco.

- Presentación del desarrollo tecnológico "Traductor Wixárika-Español" en *Séptimo Seminario de Ingeniería Lingüística* (SIL) en el Instituto de Ingieniería de la UNAM, el 9 de Septiembre de 2016.

- Artículo R. C. Barron, J. M. Mager Hois, y F. Reyes Avilés, "Richard feynman, los alfabetos y los lenguajes", *Reling Lingüistica Aplicada*, vol. 10, Junio 2016. (Barron et al. 2016)

- Artículo J. M. Mager Hois, C. Barron Romero, and I. V. Meza Ruíz, "Traductor estadístico wixárika - español usando descomposición morfológica" *COMTEL*, no. 6, Septiembre 2016.(Mager Hois et al. 2016)

El estudio en el campo de la traducción automática es dinámico, con paradigmas cambiantes que ofrecen nuevas perspectivas de estudio. Dado que para todos los mo-

[1] `http://turing.iimas.unam.mx/wix`
[2] `http://github.com/pywirrarika`

delos es el segmentador morfológico una herramienta previa a la traducción y a la alineación, será adecuado buscar mejorar su funcionamiento. Si bien el FST ha generado segmentaciones satisfactorias para el proyecto, se podrán usar otras propuestas, como es Morfessor (Grönroos et al. 2014) para segmentación semi-supervisada y no supervisada, FST probabilísticos (Sak et al. 2009) y plantear soluciones de desambiguación entre posibles segmentaciones basadas en aprendizaje máquina. Estos métodos también podrán mejorar el etiquetado de los morfemas y agregar información complementaria a la traducción.

Otro aspecto para trabajo futuro es mejorar la alineación entre palabras de GIZA++, dado que tomar los morfemas como palabras genera alienaciones no óptimas. Trabajos como los de Eyigöz (Eyigöz 2014) plantean una doble alineación, tanto de palabra a palabra, como de morfema a palabra. El trabajo sobre mejores modelos de alineación, que permitan emparejar de manera efectiva palabras y morfemas, llevaría a incrementar la calidad de traducción para lenguas aglutinantes.

En el caso de traducción con bajos recursos también existen alternativas para ampliar su estudio. Si bien, en este trabajo se utiliza un diccionario bilingüe, ha mostrado buenos resultados con el uso de diccionarios jerárquicos y expansiones con conocimiento gramatical del texto origen, mediante un análisis de posiciones sintácticas (POS), aplicando reglas predefinidas (Nießen & Ney 2004). El uso de *WordNet*, diccionarios jerárquicos y POS requieren, a su vez, la creación de los mismos, dado que no existen en este momento. Proyectos como *Verbomil* en Alemania, financiados por el gobierno, han permitido crear estos recursos para ciertas lenguas europeas que son costosas de generar por la cantidad de tiempo humano requerido. La ampliación del código también es posible a partir de técnicas de extracción de corpus paralelo, como se ha estudiado para el náhuatl por Gutiérrez (Gutierrez-Vasques 2015).

También será necesario tratar el problema del ruido de los idiomas indígenas en sus escritos. Como se ha mencionado, las diferentes escrituras, pero sobre todo, los dialectos de los mismos idiomas, presentan un reto mayor. La irregularidad de los mismos requiere formas para tratarlos de manera adecuada. Estudios del NLP, como el caso del náhuatl de Gutiérrez (Gutierrez-Vasques et al. 2016), reportan el mismo problema.

Otro trabajo futuro importante será crear traductores a más lenguas indígenas, dada la facilidad de adaptación del modelo propuesto a nuevos pares de idiomas. La mayor parte de lenguas originarias del continente americano son aglutinantes y de gran complejidad morfológica. Como se ha mencionado, el wixárika pertenece a la familia de lenguas yutonahuas, con las cuales comparte fuertes vínculos sintácticos, gramati-

cales y semánticos. También es posible incorporar interfaces de voz y tinta electrónica, que facilitarán la interacción de las personas con sistemas de traducción automática. El progreso en la tecnología de tabletas, celulares y procesadores programables hacen atractivo el diseño y la construcción de una aplicación o de un dispositivo de traducción automática personal. Este tipo de herramientas fomentan la vitalización de las lenguas originarias en un entorno marcado por las TIC.

Con la traducción automática, se podrá acercar gran cantidad de textos o información entre culturas y hablantes del wixárika y del español. Desde traducir interfaces computacionales, para permitir a los hablantes usar la tecnología en su propio idioma, hasta traducir textos literarios de las dos culturas, mejorar la defensa de los hablantes frente a los tribunales, traducir leyes y demás información oficial del Estado. A partir de la "Ley General de Derechos Lingüísticos de los Pueblos Indígenas" (Oficial 2003) todas las lenguas indígenas son reconocidas con la misma validez que el español y las instancias de gobierno deberán difundir las leyes, reglamentos y programas en estas lenguas.

Retomando un artículo escrito en el contexto de este trabajo, titulado "Richard Feynman, los alfabetos y los lenguajes" (Barron et al. 2016), cada lenguaje natural surge en un proceso cultural e histórico, adoptando una estructura para poder modelar la realidad y las ideas humanas, conservando sus matices y peculiaridades. Dado que no existe una traducción sin pérdida de información, cada vez que un lenguaje desaparece, la humanidad pierde una parte de su semántica universal. Impulsar un lenguaje por su eficiencia o facilidad de aprendizaje por sobre otros, como es el caso del inglés en nuestros días, con la idea de una equivalencia entre ellos, es un error que empobrecerá al conjunto. El modelo educativo en México además de impulsar el inglés, también debería darle un lugar importante a los idiomas de los pueblos originarios que encierran grandes riquezas. Bajo una visión errada de "lenguas inferiores", son descuidadas, discriminadas y corren el riesgo de desaparecer.

El problema de traducción sigue siendo un problema duro y la traducción automatizada es un reto aún mayor. A pesar de encontrarse muy desarrollada la MT para ciertos pares de idiomas, los resultados aún son imperfectos. La cuestión se acentúa si se considera la falta de herramientas lingüísticas computacionales para idiomas como el wixárika. El presente traductor español-wixárika abre las puertas para acercar una gran cantidad de textos a los pueblos originarios en su propia lengua, e involucrar al hablante del español a la rica semántica del wixárika.

Bibliografía

Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D. & Yamada, K. (2002*a*), 'Translation with scarce bilingual resources', *Machine Translation* **17**(1), 1–17.

Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D. & Yamada, K. (2002*b*), 'Translation with scarce bilingual resources', *Machine Translation* **17**(1), 1–17.
URL: *http://www.jstor.org/stable/40008207*

Bahattacharyya, P. (2015), *Machine Translation*, CRC Press.

Barron, R. C., Mager Hois, J. M. & Reyes Avilés, F. (2016), 'Richard feynman, los alfabetos y los lenguajes', *Relingüistica Aplicada* **10**(19).
URL: *http://relinguistica.azc.uam.mx/no019/*

Bender, O., Zens, R., Matusov, E. & Ney, H. (2004), Alignment templates: the rwth smt system, *in* 'International Workshop on Spoken Language Translation', Kyoto, Japan, pp. 79–84.

Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., Mercer, R. L., Printz, H. & Ureš, L. (1994), The candide system for machine translation, *in* 'Proceedings of the Workshop on Human Language Technology', HLT '94, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 157–162.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R. & Roossin, P. (1988), A statistical approach to language translation, *in* 'Proceedings of the 12th Conference on Computational Linguistics - Volume 1', COLING '88, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 71–76.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D. & Mercer, R. L. (1993), 'The mathematics of statistical machine translation: Parameter estimation', *Comput. Linguist.* **19**(2), 263–311.

Calderón, H. D., Mamani Calderón, C. D., Cagniy, C. & Mamani Calderón, E. F. (2009), 'Translation with scarce bilingual resources', *Revista Investigación* **5**(3).

de Lenguas Indígenas, I. N. (2016), 'Lenguas indígenas en méxico y hablantes (de 3 años y más) al 2015'.
URL: *http://cuentame.inegi.org.mx/hipertexto/todas_lenguas.htm*

Ermolaeva, M. (2014), An adaptable morphological parser for agglutinative languages, *in* 'Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014', Pisa University Press, pp. 164–168.

Eryiğit, G. & Adalı, E. (2004), An affix stripping morphological analyzer for Turkish, *in* 'Proceedings of the International Conference on Artificial Intelligence and Applications', Innsbruck, pp. 299–304.

Eyigöz, E. (2014), Morphology Modeling for Statistical Machine Translation, PhD thesis, University of Rochester, Rochester, New York.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G. & Tyers, F. M. (2011), 'Apertium: a free/open-source platform for rule-based machine translation', *Machine Translation* **25**(2), 127–144.

Geographic, N. (2016), 'El sueño de un traductor al náhuatl'.
URL: *http://www.ngenespanol.com/traveler/tecnologia/14/05/23/sueno-traductor-al-nahuatl/*

Grimes, J. E. (1964), *Huichol Sintax*, Series Practica, Mouton & Co.

Grönroos, S.-A., Virpioja, S., Smit, P. & Kurimo, M. (2014), *Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology*, Dublin City University and Association for Computational Linguistics, pp. 1177–1185. VK: triton coin.

Guillermo, B. B. (1981), *Utopia y revolución: El pensamiento político contemporáneo de los indios en America Latina*, 1 edn, Editorial nueva imágen, México.

Gutierrez-Vasques, X. (2015), Bilingual lexicon extraction for a distant language pair using a small parallel corpus, *in* 'Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop', Association for Computational Linguistics, Denver, Colorado, pp. 154–160.
URL: *http://www.aclweb.org/anthology/N15-2021*

Gutierrez-Vasques, X., Sierra, G. & Pompa, I. H. (2016), Axolotl: a web accessible parallel corpus for spanish-nahuatl, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)', European Language Resources Association (ELRA), Paris, France.

Gómez, P. (1999), *Huichol de San Andrés Cohamiata, Jalisco*, Archivo de lenguas indígenas de México, Colegio de México.

Hoang, H. & Koehn, P. (2008), Design of the moses decoder for statistical machine translation, *in* 'Software Engineering, Testing, and Quality Assurance for Natural Language Processing', SETQA-NLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 58–65.

Hopcroft, J. E., Motwani, R., Rotwani & Ullman, J. D. (2000), *Introduction to Automata Theory, Languages and Computability*, 2nd edn, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Hutchins, W. J. (2004), *The Georgetown-IBM Experiment Demonstrated in January 1954*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 102–114.

INALI (2016), 'Padrón nacional de intérpretes y traductores en lenguas indígenas'.
URL: *http://panitli.inali.gob.mx/*

Iturrio, J. L. & Gómez López, P. (1999), *Gramática Wixarika I*, Archivo de lenguas indígenas de México, Lincom Europa.

José, G. (2009), *Hablemos español y huichol*, Lingüístico Verano, Tlalpan, México.

Julio, X. (1993), *Wixárika niawarieya / La canción huichola.*, Universidad de Guadalajara, Guadalajara, México.

Jurafsky, D. & Martin, J. H. (2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edn, Prentice Hall PTR, Upper Saddle River, NJ, USA.

Knight, K. (1999), 'Decoding complexity in word-replacement translation models', *Comput. Linguist.* **25**(4), 607–615.
URL: *http://dl.acm.org/citation.cfm?id=973226.973232*

Koehn, P. (2004), *Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 115–124.

Koehn, P. (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, *in* 'Conference Proceedings: the tenth Machine Translation Summit', AAMT, AAMT, Phuket, Thailand, pp. 79–86.

Koehn, P. (2010), *Stadistical Machine Tanslation*, Cambridge University Press.

Koehn, P., Och, F. J. & Marcu, D. (2003), Statistical phrase-based translation, *in* 'Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1', NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 48–54.

Laukaitis, A. & Vasilecas, O. (2007), *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, chapter Asymmetric Hybrid Machine Translation for Languages with Scarce Resources, pp. 397–408.
URL: *http://dx.doi.org/10.1007/978-3-540-70939-8_35*

Mager Hois, J. M., Barron Romero, C. & Meza Ruíz, I. V. (2016), 'Traductor estadístico wixarika - español usando descomposición morfológica', *COMTEL* (6).

Martínez Casas, R. (2011), 'La formación de los profesionistas bilingües indígenas en el méxico contemporáneo', *Perfiles educativos* **33**, 250 – 261.

McIntoch, J. & Grimes, J. (1954), *Vocabulario huichol-castellano, castellano-huichol*, Instituto Lingüístico de Verano, D.F., México.

Microsoft (2016), 'Microsoft translator community partners'.
URL: *https://www.microsoft.com/en-us/translator/community.aspx*

Moses (2016), 'Moses statistical machine translation system'.
URL: *http://www.statmt.org/moses/*

Nießen, S. & Ney, H. (2004), 'Statistical machine translation with scarce resources using morpho-syntactic information', *Computational Linguistics* **20**(2), 181–204.

Och, F. J. (1999), An efficient method for determining bilingual word classes, *in* 'Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics', EACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 71–76.
URL: *http://dx.doi.org/10.3115/977035.977046*

Och, F. J. & Ney, H. (2000), Improved statistical alignment models, *in* 'Proceedings of the 38th Annual Meeting on Association for Computational Linguistics', ACL '00, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 440–447.
URL: *http://dx.doi.org/10.3115/1075218.1075274*

Och, F. J. & Ney, H. (2002), Discriminative training and maximum entropy models for statistical machine translation, *in* 'Proceedings of the 40th Annual Meeting on Association for Computational Linguistics', ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 295–302.

Och, F. J. & Ney, H. (2003), 'A systematic comparison of various statistical alignment models', *Computational Linguistics* **29**(1), 19–51.

Och, F. J., Tillmann, C., Ney, H. & Informatik, L. F. (1999), Improved alignment models for statistical machine translation, *in* 'University of Maryland, College Park, MD', pp. 20–28.

Oficial, D. (2003), 'Ley general de derechos lingüísticos de los pueblos indígenas'.

Oflazer, K. (2008), *Statistical Machine Translation into a Morphologically Complex Language*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 376–387.

P. E. Hart, N. J. N. & Raphael, B. (1968), 'A formal basis for the heuristic determination of minimum cost paths', *IEEE Transactions on Systems, Science, and Cybernetics* **SSC-4**(2), 100–107.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), Bleu: A method for automatic evaluation of machine translation, *in* 'Proceedings of the 40th Annual Meeting on Association for Computational Linguistics', ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 311–318.
URL: *http://dx.doi.org/10.3115/1073083.1073135*

Roy, M. (2010), Approaches to handle scarce resources for Bengali Statistical Machine Translation, PhD thesis, Simon Fraser University, Burnaby, BC, Canada.

Sak, H., Güngör, T. & Saraçlar, M. (2009), A stochastic finite-state morphological parser for turkish, *in* 'Proceedings of the ACL-IJCNLP 2009 Conference Short Papers', ACLShort '09, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 273–276.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006*a*), A study of translation edit rate with targeted human annotation, *in* 'In Proceedings of Association for Machine Translation in the Americas', pp. 223–231.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006*b*), A study of translation edit rate with targeted human annotation, *in* 'In Proceedings of Association for Machine Translation in the Americas', pp. 223–231.

Tarski, A. (1936), 'Der Wahrheitsbegriff in den formalisierten Sprachen', *Studia Philosophica* **1**, 261–405.

Tillmann, C. (2001), Word Re-Ordering and Dynamic Programming based Search Algorithm for Statistical Machine Translation, PhD thesis, Rheinishch-Westfëlishen Technishen Hochschule Aachen, Alemania.

UNESCO (2003), 'Convención para la salvaguardia del patrimonio cultural inmaterial'.
URL: *http://unesdoc.unesco.org/images/0013/001325/132540s.pdf*

UNESCO (2007), 'Elaboración de una convención para la protección de las lenguas indÍgenas y las lenguas en peligro'.
URL: *http://unesdoc.unesco.org/images/0015/001503/150360s.pdf*

Vargas, M. P. (1978), *La llave del huichol*, SEP INAH.

Wahlster, W. (1997), *VERBMOBIL: Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache*, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 215–224.

Zechner, K. & Waibel, A. (2000), Minimizing word error rate in textual summaries of spoken language, *in* 'Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference', NAACL 2000, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 186–193.
URL: *http://dl.acm.org/citation.cfm?id=974305.974330*

Zens, R., Och, F. J. & Ney, H. (2002), *Phrase-Based Statistical Machine Translation*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 18–32.

Simbología

f^J := Frase en lengua original.

e^I := Frase en lengua destino.

f'^J := Frase en lengua original descompuesta morfológicamente.

Σ := Alfabeto.

Σ^* := Cerradura de Kleene.

L := Lenguaje, tal que $L \subset \Sigma^*$.

$p(f^J|e^I)$:= Regla de bayes, probabilidad de que f^J tal que e^I

argmax := Argumentos para obtener el máximo.

$p_{LM}(e)$:= Modelo del lenguaje e.

$\phi(\bar{f}|\bar{e})$:= Tabla de traducción por frases.

d := Modelo de alineación.

Ξ := Función de descomposción morfológica.

ϵ := Palabra vacía.

$n(\phi|f)$:= Modelo de fertilidad.

A := Conjunto de alineamientos

count(...) := Contabilizar

$lev_{a,b}(a,b)$:= Distancia de Levenshtein.

Acrónimos

ATG Análisis transferencia generación.

ATS Alignment Template System.

BLEU Bilingual Evaluation Understudy.

CGM Con Segmentación Morfológica.

CSEM Con Segmentación y Etiquetado Morfológica.

EBMT Example Based Machine Translation.

FST Finite State Transducer.

GIZA++ paquete de alineado de textos, con modelos IBM.

GNU GNU is Not Unix. Sistema operativo libre, clon de Unix.

GPL General Public Licence.

IBM International Business Machines.

INEGI Instituto Nacional de Estadística, Geografía e Información.

LM Language Model (Modelo de lenguaje).

MOSES Moses translation system.

NLP Procesamiento de Lenguaje Natural, en inglés Natural Language Processing.

NP-Complete Non Polinmial-Complete.

RBMT Rule based Machine Translation.

SEP Secretaría de Educación Pública.

SGM Sin Segmentación Morfológica.

SMT Stadistic Machine Translation.

SOV Sujeto Objeto Verbo.

SVO Sujeto Verbo Objeto.

TIC Tecnologías de la Información y Comunicaciones.

TA Traducción automática.

TER Translation Error Rate.

WER Word Error Rate.

Glosario

$n$gramas Es una secuencia contigua de n elementos de una secuencia dada de texto. 1

Análisis de patrones Es una rama del aprendizaje máquina que se enfoca en el reconocimiento de regularidades en los datos. 1

Aprendizaje Máquina Rama de las ciencas de la computación que estudia los algoritmos que pueden aprender y hacer predicciones a partir de los datos. 1

Autómata de Estados Finitos modelo computacional que realiza cómputos en forma automática sobre una entrada para producir una salida. 1

Corpus paralelo Colección de texto compuesta por frases en dos idiomas que han sido producto de un proceso de traducción humana. 1

Frase Tupla de palabras que conforman una unidad de ideas. 1

Interlingua Lenguaje abstracto independiente usado para fungir como intermediario de dos o más lenguas. 1

Lengua Aglutinante Lengua en que las palabras son compuestas por diferentes morfemas que determinan su significado y que no se modifican después de la unión. 1

Lengua Polisintética Lengua donde sus palabras se componen por muchos morfemas. 1

Lenguaje natural Es una lengua que ha evolucionado naturalemnte entre los humanos por uso y repetición sin planeación alguna. 1

Minería de datos Es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que se presentan en ellos. 1

Morfema Unidad más pequeña de la lengua que tiene significado léxico o gramatical y no puede dividirse en unidades significativas menores. 1

Segmentador Morfológico Función que descompone una palabra aglutinada en los morfemas que la componen. 1

Semántica Es el estudio del significado. 1

Token Elemento básico de una frase equivalente a una palabra. 1

Traducción híbrida Combinación de paradigmas de traducción automáticos. 1

Traductor Expresar en una lengua lo que está escrito o se ha expresado antes en otra. 1

Transductor de Estados Finitos Es un Autómata de Estados Finitos deterministas con transiciones sobre parejas de símbolos. 1

Wixárika La lengua wixárika, también conocida como huichol, es una lengua indígena hablada en los estados mexicanos de Jalisco, Nayarit, Zacatecas y Durango. Tiene entre treinta y cincuenta mil hablantes, y pertenece a la familia yutoazteca. 1

Índice alfabético

Apéndice A

Código

Los siguientes programas escritos en Python 3 cubren las tareas descritas en los capítulos anteriores para el análisis de lenguaje natural para el idioma wixárika. Los presentes códigos se encuentran publicados bajo la licencia GPL versión 3[1]. Al ser libres, nuevas características se incorporarán en la versión online.

A.1. Segmentador Morfológico

El guión *wmorph.py* realiza la segmentación de una palabra wixárika aglutinada. No distingue si es aglutinable o no, simplemente intenta realizar la segmentación. Devuelve una lista de posibles firmas de segmentar la palabra. Si la lista se encuentra vacía la palabra no pudo segmentarse. El orden de las posibles segmentaciones no indica una prioridad con respecto a las demás.

En la figura 3.3 se muestra el proceso de entrenamiento. Este algoritmo es usado en el recuadro de segmentación y etiquetado. De igual forma en la figura 3.4.

Listado A.1: Segmentador mofológico

```
#!/usr/bin/env python3

import sys
import re
import codecs

#prefixes and affixes of wixarika verbs
from wixaffixes import pre, post

class Verb:
    def __init__(self, verb, debug=0):
        self.verb = verb.lower()
        #print(self.verb)
        self.paths = []
        self.roots = []
```

[1]https://github.com/pywirrarika/wixnlp

```
        self.rootslarge = []
        self.debug=debug
        Fl = codecs.open("steam.large", mode="r", encoding="utf-8")
        F = codecs.open("steam", mode="r", encoding="utf-8")
        line = F.readline()
        while 1:
            line=line.replace("\n", "")
            self.roots.append(line)
            line=F.readline()
            if not line:
                break
        if self.debug:
            print("**************************")
            print(self.roots)

        line = Fl.readline()
        while 1:
            line=line.replace("\n", "")
            self.rootslarge.append(line)
            line=Fl.readline()
            if not line:
                break
        if self.debug:
            print("**************************")
            print(self.rootslarge)


        self.start()

    def start(self, prev="", pos=0, path=[]):
        if pos > len(pre)-1:
            return
        if self.debug:
            print("New␣branch:␣", str(pos), str(prev), str(path))
        gotone = False
        for s in pre[pos]:
            s_reg = s.replace("+", "\+")
            prev_reg=prev.replace("+", "\+")
            if self.debug:
                print("Searching␣^"+prev_reg+s_reg+"+")
            reg = re.compile("^"+prev_reg+s_reg+"+")
            m = reg.match(self.verb)
            if m:
                gotone= True
                if self.debug:
                    print("Found:" + str(pos) + m.group())
                nprev = m.group()
                npath = list(path)
                npath.append(("+str(pos)+"", s))
                self.start(nprev, pos+1, npath)
                nprev = nprev.replace("+","\+")
                for root in self.roots:
```

```
                root2 = root.replace("+","\+")
                rootmatch=re.compile("^"+nprev+root2+"+")
                rm = rootmatch.match(self.verb)
                if rm:
                    if self.debug:
                        print("Found:" + "[root]" + rm.group())
                        print("Found:" + "[root]" + root)
                    nrprev = rm.group()
                    nrpath = list(npath)
                    nrpath.append(("", root))#id of steam TODO

                    if self.debug:
                        print(nrprev)
                    if len(self.verb) == len(nrprev):
                        self.paths.append(nrpath)
                    self.end(prev=nrprev,path=nrpath)
                    continue
        if not gotone:
            if pos > 17:
                return
            self.start(prev, pos+1, path)
            return
    def end(self, prev="", pos=1, path=[]):
        if pos <= 0 or pos >= len(post):
            return
        if self.debug:
            print("Actual path", prev)
        if len(prev) == len(self.verb):
            if self.debug:
                print(path)
            return
        gotone=False
        if self.debug:
            print(str(-pos), str(post[-pos]))
        for s in post[-pos]:
            if self.debug:
                print("Actual suffix:", s, "at pos", str(pos))
            s_reg = s.replace("+", "\+")
            prev_reg= prev.replace("+", "\+")
            if self.debug:
                print("Searching ^"+prev_reg+s_reg+"+")
            reg = re.compile("^"+prev_reg+s_reg+"+")
            m = reg.match(self.verb)
            if m:
                gotone= True
                if self.debug:
                    print("Found:" + str(pos) + m.group())
                nprev = m.group()
                if self.debug:
                    print("Next search:", nprev)
                npath = list(path)
                npath.append(("-"+str(pos)+"", s))
```

```
                if len(self.verb) == len(nprev):
                    self.paths.append(npath)
                self.end(nprev, pos+1, npath)
        if not gotone:
            self.end(prev, pos+1, path)

class Word:
    def __init__(self, model_file):
        F = codecs.open("dic", mode="r", encoding="utf-8")
        self.dic = {}
        self.symbols = ['U+F[illegible]FFD,].'
        self.model = io.read_binary_model_file(model_file)

        line = F.readline()
        while line:
            line = line.split()
            if line:
                self.dic[line[0]] = line
            line = F.readline()
    def checkdic(self,word):
        if word in self.dic.keys():
            try:
                pos = self.dic[word][1]
            except:
                pos =""
            print(word, end="␣")
        else:
            if word in self.symbols:
                print(word, end="␣")
            else:
                #print(word, "[Nid]", end=" ")
                seg = self.model.viterbi_segment(word)
                for affix in seg[0]:
                    print(affix, end="␣")


if __name__ == "__main__":
    if len(sys.argv) != 2:
        print("Formato:")
        print("␣␣␣␣␣wmorph.py␣word")
        sys.exit()
    v = Verb(sys.argv[1], debug=1)
    print("Found␣paths")
    print(v.paths)
```

wixaffixes.py contiene la información sobre posiciones y morfemas de los verbos aglutinados en wixárika. Esta información es usada por *wmorph.py*.

Listado A.2: Morfemas del wixárika

```
pre = [
        ["'a","a", "tsi", "ke", "u","'u", "e", "'e"], # pos 18
```

```
        ["ne", "pe", "te", "xe", "me"], #pos 17
        ["ka", "r+ka", "n+ka", "ke"], #pos 16
        ["p+", "m+", "p"], #pos 15
        ["ka"], #pos 14
        ["ka"], #pos 13
        ["ne", "ma", "ta", "xe", "r"], #pos 12
        ["ti", "tsi"], #pos 11
        ["ta", "xe"], #pos 10
        ["ni", "n"], #pos 9
        ["wa"], #pos 8
        ["'u", "u", "ha", "e","he", "heu", "eu"], #pos 7
        ["i", "'i"], #pos 6
        ["'ana", "'anu", "wa", "ana", "anu"], #pos 5
        ["ne", "a", "'a", "ta", "yu"], #pos 4
        ["ti", "ta", "ku", "ka", "ye"], #pos 3
        ["red"], #pos 2
        ["ti", "ta","ku", "ka", "ye"] #pos 1
    ]

post = [
        ["12"], # End pos
        ["kaku"], #pos 24
        ["t+"], #pos 23
        ["ka","kai"], #pos 22
        ["t+"], #pos 21
        ["ni","ka"], #pos 20
        ["t+"], #pos 19
        ["k+"], #pos 18
        ["tsie", "yari"], #pos 17
        ["kai", "yu"], #pos 16
        ["ke"], #pos 15
        ["x+a"], #pos 14
        ["kai", "tei", "x+", "ni", "m+k+"], #pos 13
        ["t+","kaku","ka","ku","me","yu"], #pos 12 !!! Terminal
        ["t+", "t+ka"], #pos 11
        ["xime","r+me","ne","t+we","we","wawe","m+k+","ku","mie","yu","n+a","ka","wa
        ["x+a"], #pos 9
        ["rie"], #pos 8
        ["ya"], #pos 7
        ["t+a","rie"], #pos 6
        ["tsi"], #pos 5
        ["ka", "ta", "ya", "rie"], #pos 4
        ["t+a","ta","ya","rie", "kie", "ke", "ma", "wie", "xie"], #pos 3
        ["ka","y+", "ya", "y+k+", "m+"], #pos 2
        ["t+"]] #pos 1
```

A.2. Análisis de texto wixárika

El programa *wixpre.py* analiza un texto íntegro en wixárika, y decide que palabras serán segmentadas y cuales no con ayuda del diccionario del apéndice B. La segmenta-

ción es realizada por *wmorph.py*.

Listado A.3: Análisis de un texto wixárika

```
#!/usr/bin/env python3

import sys
from normwix import normwix, tokenizewix
from seg import segment, segtext

sin = 0
Fo = 0

if __name__ == "__main__":

    if len(sys.argv) < 2 or len(sys.argv) > 4:
        print("Formant:")
        print("␣␣␣␣␣wixpre.py␣inputfile␣[-s]␣[outputfile]")
        sys.exit()

    if len(sys.argv) == 4:
        if sys.argv[2] == "-s":
            sin = 1
            print("Writing␣to␣", sys.argv[3], "without␣morph␣anotations")
            outfile = sys.argv[3]
            Fo = open(outfile, "w")

    elif len(sys.argv) == 3:
        print("Writing␣to␣", sys.argv[2])
        outfile = sys.argv[2]
        Fo = open(outfile, "w")

    infile = sys.argv[1]
    Fi = open(infile, "r")
    text = Fi.read()
    Fi.close()
    text = normwix(text)
    text = tokenizewix(text)
    text = segtext(text, s=sin)

    if Fo == 0:
        print(text)
    else:
        print("Writing␣to␣", sys.argv[2])
        Fo.write(text)
        Fo.close()
```

A.3. Normalizado y tokenizado

Para poder analizar de manera correcta los archivos en wixárika se unifican las posibles escrituras, y se decide que es considerado como palabra. Esta tarea es realizada por *normwix.py*.

Las tareas de normalizado y tokenizado del wixárika se utilizan tanto en el proceso de entrenamiento, como en la traducción de wixárika al español, lo cual se puede ver en las figuras 3.3 y 3.4.

Listado A.4: Normalización y tokenización

```
#!/usr/bin/env python3

import sys
import re

def normwix(text):
    text = text.lower()
    text = re.sub(r"[U+FFFD] "'", text, flags=re.IGNORECASE)
    #text = re.sub(r"'", "", text, flags=re.IGNORECASE)
    text = re.sub(r"v", "w", text, flags=re.IGNORECASE)
    text = re.sub(r"c", "k", text, flags=re.IGNORECASE)
    text = re.sub(r"[0-9]+", "", text, flags=re.IGNORECASE)
    text = re.sub(r"ch", "ts", text, flags=re.IGNORECASE)
    text = re.sub(r"rr", "x", text, flags=re.IGNORECASE)
    text = re.sub(r"␣+", "␣", text, flags=re.IGNORECASE)
    text = re.sub(r"^␣", "", text, flags=re.IGNORECASE)
    text = re.sub(r"[áàä]", "a", text, flags=re.IGNORECASE)
    text = re.sub(r"[éèë]", "e", text, flags=re.IGNORECASE)
    text = re.sub(r"[íìï]", "i", text, flags=re.IGNORECASE)
    text = re.sub(r"[óòö]", "o", text, flags=re.IGNORECASE)
    text = re.sub(r"[úùü]", "u", text, flags=re.IGNORECASE)

    #text = text.replace("á", "a")

    text = re.sub(r"([a-z])\1+", r"\1", text, flags=re.IGNORECASE)
    return text


def tokenizewix(text):
    text = re.sub(r"[^\s]([.|,|\-,\"|:|[U+FFFD][U+FFFD])",␣r" \1",␣text)
␣␣␣␣text␣=␣re.sub(r"([.|,|\-,\"|:|[U+FFFD][U+FFFD])[^\s]", r"\1␣", text)
    return text

if __name__ == "__main__":

    l = 4
    op = sys.argv[1]
    if not "-" in op:
        l = 3
        op = ""
```

```
    else:
        if "p" in op:
            l = 3
        else:
            outfile = sys.argv[3]
            Fo = open(outfile, "w")

    if len(sys.argv) != l:
        print("Formant:")
        print("␣␣␣␣␣normwix.py␣[-a|-n|-t|-p]␣inputfile␣[outputfile]")
        sys.exit()

    infile = sys.argv[2]
    Fi = open(infile, "r")
    text = Fi.read()
    Fi.close()
    if ("n" in op) or ("a" in op):
        text = normwix(text)
    if ("t" in op) or ("a" in op):
        text = tokenizewix(text)
    if "p" in op:
        print(text)
    else:
        Fo.write(text)
        Fo.close()
```

A.4. Identificación de raíces

A partir de un texto wixárika segmentado manualmente, se extraen las raíces identificadas en ese corpus y se genera un archivo donde serán usadas por *wmorph.py*.

Listado A.5: Extracción de raíces y palabras no aglutinadas

```
# -*- encoding:utf-8 -*-
#!/usr/bin/env python3

import sys
from wixaffixes import pre, post

def getmorphpre(verb, debug=0):
    level=0
    seglist = []
    preval = 1
    steam = ""
    for morph in verb:
        if debug:
            print(morph)
        nogot = 1
        if preval:
            if debug:
```

```
                print("Prefix...")

            while nogot:
                if level > 17:
                    if debug:
                        print("Level␣17!")

                    e = morph+""
                    seglist.append(e)
                    level=0
                    preval=0
                    nogot=0
                    steam=morph
                    continue
                if morph in pre[level]:
                    if debug:
                        print(morph, "in␣level", str(level))
                    e = morph+""+str(level)+""
                    seglist.append(e)
                    level = level+1
                    if debug:
                        print(seglist)
                    nogot = 0
                else:
                    level = level+1
        else:
            while nogot:
                if level > 22:
                    if debug:
                        print(seglist, end="␣")
                    else:
                        for m in seglist:
                            print(m, end="␣")
                    return
                if morph in post[level]:
                    e = morph+""+str(-1*level)+""
                    seglist.append(e)
                    level = level+1
                    nogot = 0
                else:
                    level = level+1
    if len(verb) != len(seglist) or not steam:
        if debug:
            print("(ERROR){{", verb, str(seglist), end="}}␣")
        else:
            for m in verb:
                print(m, end="␣")
    else:
        if debug:
            print(seglist, end="␣")
        else:
            for m in seglist:
```

```
                print(m, end="␣")
    return [steam, seglist]

def get():
    F = open("train.wix", "r")
    line = F.readline()
    splitdoc = []
    noverbs = []
    steams = []
    while line:
        wordsep = []
        sep = line.split()
        for word in sep:
            wsep = word.split("-")
            wordsep.append(wsep)
        splitdoc.append(wordsep)
        line=F.readline()
    for line in splitdoc:
        for word in line:
            if len(word) == 1:
                print(word[0], end="␣")
                noverbs.append(word[0])
            else:
                res = getmorphpre(word)
                try:
                    steams.append(res[0])
                except:
                    pass
        print()
    F.close()
    noverbsset = list(set(noverbs))
    steamsset = list(set(steams))
    #print(str(len(noverbsset)), "No verbs:", str(noverbsset))
    #print(str(len(steamsset)), "Steams:", str(steamsset))

    Fdic = open("dic", "r")
    Fsteams = open("steams.txt", "r")

    fsteams = Fsteams.read().split('\n')
    st = [s.split("=")[0] for s in fsteams]
    #print(st)
    [st.append(ns) for ns in steamsset]
    stf = set(st)
    #print(str(len(stf)), "Steams:", str(stf))
    newsteamfile = open("steam2", "w")
    [newsteamfile.write(line+"\n") for line in stf]
    newdicfile = open("dic2", "w")
    [newdicfile.write(w+"\n") for w in noverbsset]


    newsteamfile.close()
    Fdic.close()
```

```
    Fsteams.close()


if __name__ == "__main__":
    if len(sys.argv) != 2:
        get()
        sys.exit(1)
    verb = sys.argv[1].split("-")
    getmorphpre(verb, debug=1)
```

A.5. Identificación de texto wixárika

El guión *idtexto.py* identifica un texto wixárika que se encuentra incrustado dentro de frases en otro idioma, al igual que el texto completo en wixárika. Este guión es utilizado para aumentar el corpus.

Listado A.6: Identificación de texto wixárika

```
#Copyright (c) Jesús Manuel Mager Hois 2016
#
#Permission is hereby granted, free of charge, to any person obtaining a
#copy of this software and associated documentation files (the
#    "Software"), to deal in the Software without restriction, including
#without limitation the rights to use, copy, modify, merge, publish,
#distribute, sublicense, and/or sell copies of the Software, and to
#permit persons to whom the Software is furnished to do so, subject to
#the following conditions:
#
#The above copyright notice and this permission notice shall be included
#in all copies or substantial portions of the Software.
#
#THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS
#OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF
#MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT.
#IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY
#CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT,
#TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE
#SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.


import sys
import pickle as pk
import sys

import nltk
from nltk import word_tokenize
from nltk.util import ngrams

import wixanlp as wixa
import numpy as np
```

```
def eval_pair(lm, pair):
    """ Get probability of a pair of words """
    try:
        p = lm[pair[0]][pair[1]]
    except KeyError:
        p = 0
    return p

def eval_text(lm, text):
    tokens = word_tokenize(text)
    chain = []
    for word in tokens:
        w = list(word.lower())

        # Dividing in 2-grams
        bgs = ngrams(w, 4)
        w = []
        for bg in bgs:
            w.append(bg[0]+bg[1])
            w.append(bg[2]+bg[3])

        p = 0
        i = 0
        for i in range(len(w)-1):
            p = float(p) + eval_pair(lm, (w[i], w[i+1]))
        p = float(p) / float(i+1)

        if word in confusion:
            p = 0

        #If the word exists in the common word list, then p = .5
        # .5 is an arbitrary value
        if word in common:
            p = 0.2
        chain.append((p, word))
    return chain

if __name__ == "__main__":

    if len(sys.argv) != 2:
        print("Usage: python3 idtext.py filename.txt")
        sys.exit(1)
    filename = sys.argv[1]
    print(filename)

    confusion = 0
    try:
        f=open("confusion.pickle", "rb")
        confusion = pk.load(f)
    except OSError:
        pass
```

```
# Load the common word list
with open("common.pickle", "rb") as f:
    common = pk.load(f)
# Load the language model
with open("lm.pickle", "rb") as f:
    lm = pk.load(f)


txt = open(filename, 'r')
txt = txt.read()
txt = wixa.norm(txt, prepare=0)
chain = eval_text(lm, txt)
now = 0
lastn = 0
nextn = 2
first = 0
total = len(chain)
words = []
tmp = []
threshold = 0.01

while now < total-2:
    now += 1
    if chain[now-1][0] > threshold:
        lastn = 1
    else:
        lastn = 0
    if chain[now+1][0] > threshold:
        nextn = 1
    else:
        nextn = 0

    if (chain[now][0] > threshold) and nextn and lastn:
        if not first:
            tmp.append(chain[now - 1])
            tmp.append(chain[now])
            tmp.append(chain[now + 1])
            first = 1
        else:
            tmp.append(chain[now + 1])
    elif first:
        tmp.append(chain[now + 1])
        first = 0
        tmp2 = []
        for wt in tmp:
            if wt[0] > threshold and wt[0]:
                tmp2.append(wt)
        words.append((list(tmp2)))
        tmp = []
```

```
i = 0
j = 0
inphrase = 0

for p in words:
    for w in p:
        print(w[1], end="␣")
    print("")

print("Total␣number␣of␣phrases:", str(len(words)))
```

Apéndice B

Vocabulario wixárika-español

Para la traducción automática se auxilió del "vocabulario huichol-castellano, castellano-huichol" escrito por Grimes y McIntoch (McIntoch & Grimes 1954) en 1954. Dado que el texto encontrado fue digitalizado a partir del folleto original[1] y se tuvo que adaptar a la escritura moderna del wixárika, con ayuda del normalizador. Posteriormente fue presentado con un hablante del wixárika[2] quien realizó correcciones a la parte wixárika. Las palabras wixáritari que son sustantivos se identifican por la señalización $[s]$, seguido del morfema que se agrega para su plural. En total se tienen 1455 palabras wixáritari. Los sustantivos que no tengan un plural señalado son irregulares. El wixárika tiene una compleja forma de construir plurales, por lo que es importante agregar esta referencia.

Con el presente diccionario, hay que recordar que el proceso de traducción no es lineal y menos en una lengua polisintética como el wixárika. El diccionario se agrega como una fuente adicional de información para el modelo de traducción estadística por frases. La forma en la cual se construyen palabras en wixárika hace difícil la creación de un diccionario de todas las palabras del wixárika al español. Esto se incrementa por la posibilidad de introducir matices e información que no pueden ser contenidos en el español.

Se ha generado una lista de raíces verbales aprendidas a través del análisis morfológico de las palabras aglutinadas. En el análisis de frases, se utilizó el presente diccionario para identificar palabras que no deben ser segmentadas, y que a su vez, en el caso de los sustantivos, pueden ser utilizados como raíces verbales, adicionales a las encontradas de manera automática.

[1] Transcrito por Rebeca Guerrero Islas además de correcciones al español.

[2] Diónico Carrillo González de la comunidad de Zoquipan, Nayarit

Tabla B.1: Vocabulario

wixárika	**español**
kakái [s] te	huaraches
kakáixi [s]	especie de avispa
tetsu +akaxayari [s]	tamal de esquite
kaka+yari [s] xi	dios
kaka+yari [s] ta	lugar de adoración
kakúni [s] te	caja
kakúni [s] te	cajón
kakúni [s] te	medida para maíz
xayé [s]	cascabel
káitsa [s]	sonaja
kaiwamá	en la falda del cerro
yerí [s] te	camote de castilla
kamixa [s] te	camisa
ka'í	¡tenga!
kam+	¡mire!
kaná [s] te	frente
ter+ká ka'kuiniya[s] xi	alacrán grande que no es venenoso
tuka [s] xi	tarántula
h+ayamé [s]	cinturón
kanari [s] te	guitarra
kanéera [s]	canela
kanuwa [s] te	canoa
kanuwa [s] te	lancha
kanuwa [s] te	barco
kanuwate muwa-ye'axet+katsie [s] te	embarcación
kapé [s]	café
tsipu [s] te	chivo
kepútsa [s] te	talón
heiwat+	tal vez
deyurí	seguro que
kam+ts+ ti núani	tal vez va a venir
karará [s]	cáscara
xaweri wewiyakame [s]	instrumento de música hecho con fémur de venado
yurí	seguro
deyurí	de veras
karíma	fuerte
kwi	recio
kariú [s] te	nogal
kariúxa [s] te	cedro

wixárika	**español**
karú [s] te	plátano
'awá karú	plátano grande
m+kíte wa karú	plátano de muerto
m+kí karúya	platanillo
tetsú tukarieya [s]	fiesta de tamales de maíz
karím+xí [s] te	pestañas
+xá [s]	zacate
kar+wunátu [s]	carbonato
kákaiyari kwiniya wiwiekamé [s]	dios que enferma a los niños
katáixa [s]	guinole
atakwai ya ukwi [s] xi	lagartija que vive en las piedras
kat+ra [s] te	vela
katútsi [s] te	nuca
káunári [s] te	soga
káunári [s] te	mecate
káuxai [s] tsi	zorra
mawiyá ka-ka+yarika murata xuiwenik+ [s]	ofrenda de tejido a un dios para el éxito en tejer
kauyúmari [s]	dios mensajero entre los dioses y el cantador
kawáaya [s] tsixi	caballo
papawarí [s] rí	padrastro
haxí [s]	guaje
witaritá [s]	tiempo de aguas
kúu [s] ter+xi	víbora
kúu [s] ter+xi	reptiles
xaip+ [s] ter+xi	insectos
kúu teuta mieme	víbora ponzoñosa
kuáits+ [s] te	anzuelo
kuámu [s] xi	faisán
its+ [s]	bastón
kuar+pá [s] te	ciruela
kuatsá [s] ri	cuervo
kuatsápa+ [s] te	cadera
nawi tumini [s]	setenta y cinco centavos
kuatá [s] te	guaiz
kuatá [s] te	guache colorado
maxá tsimupé [s]	venado chico
kuatemukáme [s]	cervatillo
kuatú [s] xi	bola negra

wixárika	español
kuatú [s] xi	fruto
kuá'uni [s] te	humo
kuáuxá [s] te	olote
kuauyáarí [s] te	mazorca
ke'uxá [s] te	quelite
kuáaxa [s] te	higuerilla
kuaxí [s] te	cola de animal
kuaxíya [s]	sudor
kuaxú [s] xi	garza
kuaxúa [s] te	sesos
kúuka [s] te	chaquira
kúuka [s] te	cuello de chaquíra
kukúri [s] te	chile
kúukurú [s] xi	paloma
ha+m+ [s] xi	pichón
ha+t+xí [s] xi	huilota
kemá [s] ma	cuñado
kué [s] ma	cuñada
kué [s] ma	concuña
kué [s] ma	concuño
kerí [s] xi	guajiniquil
xim+ xixí	por favor
teukarí [s] ma	abuelo materno
kuetsúkua [s] te	pegadure silvestre
kuetsúkua [s] te	pegamento
kuí	recio
kuít+	en seguida
kwi	recio
kuit+ wa	en seguida
kwi	recio
kuit+wa	en seguida
kuíni míeme	recio
kuíni míeme	en seguida
kwie [s] ta	tierra
kwiepa [s] te	suelo
kwie [s] te	terreno
kwiemúxa [s] te	algodón
kwiepúxai [s] te	malacate para hilar
kwikári [s]	canción
kwinuri [s] te	tripas
kwinuri [s] te	intestinos
kwitsári [s] te	pozole
kwitsí [s] téri	gusano
kwitá [s] te	excremento
kwitápi [s] xi	chachalaca
kwitápi [s] xi	chichalaca

wixárika	español
kwitáaxi [s] te	correa
kwitem+ [s]	lombriz
kwitemúxi [s]	lombrices
k+yé [s] teukí	palo para objetos ceremoniales
kuix+ [s] tsi	gavilán de cola roja
kuráru [s] te	corral
kurí [s] ma	hermana mayor
kutsára [s]	cuchara
kutsára+ [s]	pila sagrada de un lugar de adoración
kutsíra [s] te	machete
kutsi [s] ma	abuela
kutsikamé [s] ma	tía abuela
kweya teukari [s] ma	cuñada del abuelo
kweya kutsi [s] ma	cuñada de la abuela
kútsi [s] ma	abuela
kútsikáme[s] ma	tía abuela
teukari kweya[s] ma	cuñada del abuelo
kútsi kweya[s] ma	cuñada de la abuela
k+pi uayeyá [s] xi	capullo de mariposa cuatro espejos
kutsiyáari [s] ma	dueño
kukíya [s]	tos
kutsí [s]	sueño
ariwatsiní [s]	mensajero
n+'arika [s] ts+xi	Correo
nemetse n+'airí raye'ukamé [s]	te mando con un mensaje
tíxa+ kúxi	todavía no
kuká tsaiyakamé [s] te	collar tejido
kuxitáari [s] te	costal
kuyá [s] xi	soldado
mieriká [s] xi	guerra
mieriýa[s] xi	revolución
kuyéikame [s] te	extranjero
iwatsiká [s] te	visita
xe+matiwamé [s] te	extraño
tieriwamé	exclamación
x+ipít+a [s] te	pezcueso
k+ipí [s] te	buche
k++mána k+[s]	por medio de
k++mána [s]	por sí
k++méme [s] te	tijeras
k+ná [s] ma	esposo

wixárika	español
k++pá [s] te	cabello
k++páixá [s] te	cabello de elote
k+pí [s] xi	mariposa
k++púri [s]	cosa con textura de fibra o pelo
kaka+yari kiekamé an+ta+ye [s]	dios que vive en el norte
k+rapúxi [s] te	clavo
k+rapúxi m+'+a [s] te	clavo de olor
pirik+xá [s] te	flauta
nanayari xutsí [s]	guía de calabaza
xutsí kuapuyari [s]	tello de calabaza
xei waxawí [s] te	sien
+yakamé[s] te	mastoide
k+rípu [s] xi	concha
metsayarí mayú [s]	el mes de Mayo
itseweme [s] te	chinche de metal
meripaimiemé [s]	antepasado
k+tsi [s]	humo
tsinú [s] rixi	perrito
k+tsitúiyari [s]	ventarón
wiyeri temat+ [s]	tormenta fuerte con granizo
k+tsiúri [s] te	talega
k+tsiúri [s] te	bolsa
k++tsúnu [s] te	tronco
k++wéri [s]	lechuguilla
k+xau [s] te	tostado
k+xau watuxá [s] te	totopo hecho con sal
k+xáuri [s] te	bule
xukúri [s] te	jícara
tupiriyá [s] xi	árbol
k+yé [s] xi	palo
k+yé naiwamé [s] xi	leña
k+tsá [s] te	nalga
k+tsa [s] te	glúteo
mik+ri [s] xi	tecolote prieto
mik+ri	tecolote negro
ma+y+xa+yé [s] xi	
háa [s] te	agua
haká [s]	hambre
háka [s] te	carrizo
hakayári [s]	calabaza floreciente
hakú[s] te	otate

wixárika	español
hakuáari [s] te	tapexte
hakuáari [s] te	trampa para pescado
hakukúri [s] te	chile ancho
hakukúri [s] te	chilacate
hakuíeka [s]	dios que vive en el mar y desintegra la tierra en el tiempo de aguas
hái [s] te	nube
há+tsi [s]	sereno
há+tsi [s]	rocío
háika [k]	tres
háik+ [s]	especie de víbora
haikíri [s]	remolino chico
háitsi [s] te	tempixti
háitsi [s] te	fruta
háits+ [s] ríxi	tejón
haiw+tíri [s] te	banco de nube
haiyá [s]	hichazón
háixa [s]	ojo de agua
háa'unári [s]	ojo de agua en la tierra del peyote
hamatíana [s]	con él
hámui [s]	mezcla para hacer tejuino
hamuítsi [s] te	atole
hamúuxa [s] te	corriente del río
hapáni [s] te	tazaquillo
hapáni [s] te	tallo que da fruta como la pithaya
hap++tsáme [s]	roceador
kakaiyari kiekame teutá [s]	dios que vive en un peñasco
hapuri [s]	especie de gavilán
hakéri [s] ts+xi	niño que participa en una fiesta
haxía [s]	enojo
hakéwats+	no hay nadie
hakéwats+	no hay nada
hakíya [s]	el mes de Marzo
hakíya [s]	Cuaresma
harakúuna [s] te	laguna
haramará [s] te	mar
haramára [s]	dios del mar
haráaru [s] te	arado

wixárika	español
harawéri [s]	huerta
ite+tsiyá [s]	jardín
harayúawime [s] te	pila
harayúawimeta [s] te	agua estancada
aik+ [s] ts+xi	arriero
metserí [s]	nombre ceremonial de la luna
hariúki [s]	zacatón que se utiliza para amacizar la ranura de las flechas
har+ka [s] xi	perro de agua
hatsa [s] te	hacha
piratá [s] te	pila
haixá [s] te	ojo de agua
wiexú [s]	víbora que vive en los techos
im+ari [s] xi	semilla
tixa+	nada
naika anet+ [s] xi	rata grande
hatsukari [s] te	azúcar
tupiriya mieme aki'+t+má [s]	árbol de los arroyos
hatuxame [s] te	río
hat+a [s] te	río
hat+we [s] xi	tigrillo
katira [s] te	vela ceremonial
tau [s] kate	nombre ceremonial del sol
hawim+tari [s]	dios que vive cerca de Santa Catarina, Jalisco
kwitsi tik+memé [s] xi	gusano azotador
há+ri [s] te	capomo
há+tsi k+puri [s]	dios que vive en un peñasco
háxi [s] te	guache
háaxi [s] tsi	caimán
haxú [s] te	lodo
haxú [s] te	barro
hayú [s] xi	arbejón
hayuxari [s] xi	objeto ceremonial de barro
hayáari [s]	vasca
hakíya [s]	ayuno
hek+aríya [s]	luz

wixárika	español
heimána [s]	sobre
heimána [s]	encima de él
heimána [s]	arriba de él
hein+tsika [s]	sueño
heitsérie [s]	derecho
héiwa	un día
heiwáka	una vez
hepána [s]	hacia él
hepá+na [s]	parecido a él
hepá+na [s]	como él
hets+ana [s]	donde está él
hets+ana [s]	con él
het+ana [s]	debajo de él
he'éiyá [s]	encargo
kiekaritamé [s] xi	raza que habitaba el mundo antes de los huicholes
hikúrí [s] te	peyote
hik+	ahora
hik+	hoy
hik+pai	hasta ahora
ik+rí [s]	elote
h+pát+ [k]	otros
hípát+ [k]	los demás
h+weríka [s]	tristeza
hítua [s]	nido
hiwáatsixá [s]	fiesta de la siembra, ultima del ciclo anual, celebrada en Junio
hixitaixa [s]	maíz jiloteando
hixitarixa [s]	maíz jiloteando
hix+apa [s]	al centro
hix+ata [s]	en medio
hix+atapa [s]	al oriente
húra	cerca
huráwa	cerca
hu	sí
huriepa [s] te	estómago
huká [s] te	panza
hukátsie [s] te	barriga
hukú [s] te	pino
hukúri [s] xi	gavilán
húuna [s] ri	jején
huriekame [s]	especie de víbora
huriépa [s]	estómago

wixárika	español
húutse [s]	oso
a+raxá [s]	vinagrillo
húuta [k]	dos
hutapáari [s]	tapanco
huwíri [s]	biznaga
huxari [s] te	vellos
huxari [s] te	pelos del cuerpo
huye [s] te	camino
h+nári [s]	órgano masculino
h+nári [s]	pene
h+ri [s] te	cordón de la sierra
h+r+pána [s]	a vista de él
h+xi [s] te	ojo
h+xiéna [s]	delante de él
t+xi[s]	objetos ceremoniales de masa
máaku [s] te	mango
makúu [s] te	calabaza de castilla
makúutsi [s] te	marihuana
mái [s] te	maguey
mái [s] te	mezcal
máixa [s] te	ixtle
máixa [s] te	textil de ixtle
mamá [s] te	brazo
mamá [s] te	mano
mána	allí
manétsiki [s]	parche
maníwe [s] te	mano del metate
mara'akáme [s]	cantador
mara'akáte [s]	cantadores
maráika [s]	aura
marimá [s]	cuidado
mariutsíka	asustar
mariutsíka	espantar
maríi [s]	pájaro costeño
matsáwe [s]	especie de maguey con espinas
patsika	en cambio
wa+ka	mas
matsi [s] ma	hermano mayor
matsik+i [s]	escoba
matsikíi [s]	zacate para hacer escoba
máatsu [s] ma	sobrino
máatsu [s] ma	sobrina del hombre
matsúxí [s]	sobrino del hombre

wixárika	español
mats+wa [s] te	pulsera
mats+wa [s] te	objeto ceremonial similar a pulsera
mata [s] te	metate
matá tsikéeme [s] te	escobeta para metate
matáika [s] xi	pata de res
matáika [s] xi	lagartija
mat+ari [s]	principio
mat+ari [s]	primero
mawárí [s]	ofrenda
mawári [s]	sacrificio
máxa [s] tsi	venado
máxa kuaxí [s]	dios que vive en el oriente
máaye [s] tsi	león
ma'+ [s] ma	nieto de la mujer
ma'+ [s] ma	nieta de la mujer
meki [s] te	mezquite
merí	temprano
merí	primero
méripaí	antes
méripai	anteriormente
mérik+ts+	pues
mérik+ts+	en cuentos
merik+te	pues
merik+te	en conversación
mer+kaxa [s] te	aguacatillo
mer+karíya [s]	relámpago
mimierika [s]	rayo
metséri [s]	luna
meta [s] ri	mapache
mexikuxi	mientras que
mex+íwa	recio
mex+íwa	fuerte
mex+íma	recio
mex+íma	fuerte
mik+rí [s] xi	tecolote
mitsú [s] ri	gato
muta [s] ma	hermana menor
m+tari [s] ma	nieta del hombre
mitari [s] ma	abuelo de la mujer
m+kuá [s]	regalo
mikieriká [s]	regalo
kuruná[s]	corona
múume [s] te	frijol

wixárika	español
mumé [s] te	riñón
mumuxí [s] te	bordón
múune [s] ma	suegro
múune [s] ma	yerno del hombre
munewari [s] ma	padrastro de la esposa
niwewarí [s] ma	hijastro
muritari [s] te	cojín
muritarí [s] te	almohada
murúni [s] xi	melón
múuta [s] ma	hermano menor
muwa	allí
muwíeri [s]	plumas del cantador
muxá [s] tsi	borrego
muxá [s] tsi	cordero
maixéka [s] xi	cienpies
muxúri [s] te	guamúchil
mu'e [s]	suegro
mu'e [s]	suegra de mujer
mu'e [s]	nuera
mu'ewáari [s] ma	padrastro del esposo
mu'ewáari [s] ma	madrastra del esposo
mu'ewáari [s] ma	hijastra
mu'ú [s] te	cabeza
kimutsitá [s] te	caballete de casa
m+k+h+a [s] te	apellido
m+axá [s] te	hoja de mazorca
m+k+	aquel
m+k+mexí p	aquellos
wa+kawate [k]	muchos
m+ixa	muchos días
wa+kamex+a	muchas veces
m+pa+	así
m+kí [s] te	muerto
m+kí [s] te	difunto
m++kí mu'úya	calavera
m++ráka [s] tsíxi [s]	avispa
tutsikamé [s] ma	bisabuelo
tutsi [s] a	biznieto
meripait+ [s] a	antepasado
hairieka tsie [s] a	descendiente del tercer grado o más
m+xí [s] tsi	bagre
m+xiya [s] te	barba
m+xiya [s] te	bigote

wixárika	español
m+pa+	así
m++kiyáarí [s]	muerte
naká [s] te	oreja
nakári [s] te	nopal
nakatú [s] ts+xi	sordo
nakawe [s]	deidad que vive en 'Aitsárie
nakawe 'iwipáame [s] te	palo espinoso
maku hamuitsiyarí [s]	atole agrio de calabaza
hamuitsiyari xutsi utsiwikamé [s]	atole agrio de calabaza
nak+tsá [s]	arete
xarí [s]	olla
náime [k]	todo
náika [s] tsi	ratón
naipári [s] te	hombros
naipári [s] te	paleta
naipári [s] te	omóplato
naitsárieet+ [k]	dondequiera
naitsárie [k]	en todas partes
naitsáta [k]	a ambos lados
naitsáta [k]	a todos lados
náma [s]	objeto ceremonial tejido
náana [s]	madrina
tanáana	la Virgen
naná [s] te	raíz
naná [s] te	guía
naná [s] te	tallo
nanáwata [s]	raza mitológica
narakáxi [s]	naranja
naríka [s] te	camichín
natsatsatsa [s]	sonido que hace un perro con los dientes
náuka [k]	cuatro
nawá [s]	tejuino
nawá [s]	tesgüino
nawáaxa [s] te	cuchillo
nawí [s] te	cuero
nawí [s] te	piel
nawí [s] te	piel humana
nawí tumíni [s]	cincuenta centavos
naxí [s]	cal
naxí [s]	ceniza

wixárika	español
naxiw+yári [s]	fiesta de Febrero cuando se echa ceniza a los productos de la tierra
né	yo
néma [s] te	hígado
není [s] te	lengua
nenewíeri [s]	oración
ne+ki [s] ma	sobrino de la esposa
ne+ki [s] ma	sobrina de la esposa
ne+ki [s] ma	tío político
ne+ki [s] ma	concuño
ne+kixíwi [s] ma	sobrino de la esposa
ne+kixíwi [s] ma	sobrina de la esposa
ne+kixíwi [s] ma	tío político
ne+kixíwi [s] ma	concuño
nieríka [s] te	cara
nieríka [s] te	máscara
nieríka [s] te	disco ceremonial
nieríka [s] te	parte sagrada del coamil
niúkiperái [s] xi	hablador
niúki [s] te	palabra
niúki [s] te	idioma
níwa	vente
niwé [s] ma	hija
niwé [s] ma	hijo
matsu [s] ma	sobrino
matsú [s] ma	sobrina
matsukame [s] ma	hijo de un primo
matsukamé [s] ma	hija de un primo
matsukame [s] ma	hijo de una prima
matsukamé [s] ma	hija de una prima
niweyáame [s]	vientre
niweyáame [s]	útero
niweríkate [s]	la Campana
niweyameté [s]	constelación de estrellas
niwetsí [s] ma	sobrino de la mujer
niwetsí [s] ma	sobrina de la mujer
niwetsíka [s]	mazorcas que se guardan para la fiesta de la siembra
utaparí [s]	tapanco
niwetári [s]	tepexte
niwetári [s]	sobresaliente
niwewáari [s] ma	hijastro

wixárika	español
nu'áya [s]	hijo de él
nu'áya [s]	hija de él
nu'áya [s]	hija de ella
nu'áya [s]	hijo de ella
nuiwári [s] te	nacimiento
nuiwári [s] te	lugar de nacimiento
iwamarixí [s] te	familia
iwa [s] te	parientes
nunúutsi [s]	niño
t+rí [s]	niños
nutui [s] te	huérfano
n+'áari [s] xi	mensajero
n+'áari [s] xi	arácnido parecido al alcrán
n+'ariwáme [s]	dios del oriente que vive en 'Aitsárie
pá [s]	pan
papá [s] te	toritilla
párikuta [s]	dios que vive en la tierra del peyote
paríyatsíe [s]	tierra del peyote, San Luis Potosí
w+rikúta [s]	tierra del peyote, San Luis Potosí
patéyu [s] te	batea
páatu [s] xi	pato
pa+ri [s] te	cosa
per+ku [s] tsixi	perico
pér+	pero
pexúri [s]	pinole
p+ni [s]	posesiones
p+ni [s]	tiliches
piní [s] te	higuera
piní [s] te	amante
píx+x+i [s] tsi	pollito
turuu [s] ts+xi	buey
wakiya [s] ts+xi	novillo
putsi [s]	cosa mocha
kurítsi [s]	pene de niño
purútsa [s] te	bolsa
púti [s[te	bote
púti [s[te	lata
titá	¿qué?
képa+	¿cómo?
ké'ts+	lo que
kéipari [s]	empeine

wixárika	español
keitsariwáme [s] te	antepasado
kemári [s]	ropa
kéma [s] ma	cuñado del hombre
kemátsi [s] ma	padre del hombre
ké'ané	¿quién?
ke'ane	¿quién?–
keháate	¿quiénes?
kéepa+	¿cómo?
kepa+ka	¿cuándo?
kepútsa [s] te	talón del pie
ketsé [s] te	iguana
kets+ [s] te	pescado
ketá [s] te	pie
keurúwi [s] te	tijera del techo
kewimúta [s]	dios que vive en el mar y manda lluvias en el tiempo seco
kéexiu [s]	queso
kíi [s] te	casa
kitsie [s] te	techo
kiekáme [s]	ciudadano
kiekáme [s]	persona de la casa
kiekáme [s]	persona del rancho
kiekáme [s]	persona del país
kiekátari [s]	ciudadanos
kiekátari [s]	personas de la casa
kiekátari [s]	personas del rancho
kiekátari [s]	personas del país
kiekári [s] te	rancho
kiekári [s] te	pueblo
kiekári [s] te	ciudad
kíewíxa [s]	tiempo de lluvia
kiríwa [s] te	petaca
kir+wa [s] te	canasta grande para piscar
k+tsitáme [s] te	golondrina
k+ténie [s] te	espacio para la puerta
kie'uxa [s] te	quelite
ti'+t+wame [s]	lápiz
rétsi [s]	leche
ríma [s]	lima para afilar
tsakaimúuka [s]	dios que vive en La Mesa del Nayar, Nayarit

wixárika	español
tsakuéni [s]	raíz silvestre
tsái [s] te	sotol
tsái [s] te	mezcal para hacer agua ardiente
tsam+ráawe [s]	pluma ceremonial
tsáapa [s] ri	mojarra
tsapú [s] te	zapote
tsar+ [s] xi	hormiga arriera
tsariká [s] te	trenzas
tsáatu [s] rí	santo
tsáatu [s] rí	la Virgen
tsauréepa [s]	lugar en el centro de un templo
tsaurixíka [s]	cantador
tsaurixíka [s]	sacerdote huichol
tsa'+xi [s] ma	padre del yerno
tsa'+xi [s] ma	madre del yerno
tsa'+xi [s] ma	padre de la nuera
tsa'+xi [s] ma	madre de la nuera
tsek+xi	seguro
tsérietána [s]	a su derecha
tsikári [s]	espina de maguey
tsikwai [s] te	arrayán
tsikuáaki [s]	payaso ceremonial
ts+kuixa [s]	palo verde
tsikuráati [s] te	chocolate
tsikúri [s] te	codo
tsikúri [s] te	ángulo
tsikuwéeta [s]	escobeta
tsikuwéeta [s]	cepillo
tsikeme [s]	peine
tsik+iwíti [s] te	chiquihuite
tsik+iwíti [s] te	canasta ancho
tsik+r+i [s]	círculo
tsik+rí [s]	objeto ceremonial
tsiemp+re	siempre
tsimá [s]	amole
tsimanixi [s]	la Cabrilla
tsimanixi [s]	las siete estrellas
tsimúaka [s] r+xi	ardilla
tsim+ní [s]	órgano
mará [s]	pitahaya
tsinakari [s]	limón
ts+nari [s]	chicuatol
ts+nari [s]	atole agrio de maíz
tsinu [s] te	peine

wixárika	español
tsinú [s] r+xi	perrito
k+ts+nú [s] r+xi	perrito
tsinúxi [s]	plátano chino
tsipu [s] ri	chivo
+ráwe awakamé [s]	venado con cuernos
tsip+riki [s] te	chispa de fuego
tsip+riya [s] te	salpicadura
tsikéru [s] ts+xi	becerro
tsíki [s]	especie de chapulín
tsítsi [s] te	seno
tsítsi [s] te	teta
tawitsie [s] te	pecho
tsítsi [s] te	ubre
tsíta [s] te	cucuvixte
tsinarixa [s] te	fruta ácida
tsi+ríka [s]	hiel
tsi+ríka [s]	vesícula biliar
tsúira [s] xi	taco grueso
tsúira [s] xi	gorda
tsúira [s] xi	la última tortilla
tsumé [s]	moco
tekutsuná [s]	molcajete
tsuniya [s]	gota
tsuniriya [s]	gota
tsukiya [s]	soga de cuero crudo
tsuráakai [s] xi	pitorreal
tsuráakai [s] xi	pájaro carpintero
tsúuri [s] te	nariz
tsutíapai [s]	poniente
tsutíapai [s]	oeste
tsuwíri [s]	biznaga
tsuye [s] xi	especie de chapulín
katsuwera [s] te	casuela
ts+k+ [s]	perro
ts+k+ [s]	ceja
tsikíri [s]	perros
ts+k+ [s]	cejas
ts+rik+te	seguro
ts+rik+te	cierto
tća [s] ri	bola
tća [s] ri	fruto
tákai	ayer
taká+ [s]	nombre del sol
taka+yá [s]	nombre del sol
taka+yáatsi [s]	nombre del sol
takuá	afuera

wixárika	español
tak+ [s] te	palma
tái [s] te	lumbre
tái [s] te	fuego
táikái [s]	la tarde
tamáamata	diez
táme	nosotros
tamé [s] te	diente
tamé [s] te	ranura de flecha
táapa	al otro lado
tapiya [s]	nudo
táapikukuwi [s]	especie de pájaro chico
tárik+xa [s] te	garganta
tárik+xa [s] te	tráquea
táru [s] ma	hermano menor
táru [s] ma	hermana menor
táatu [s] te	jazmín
tatsi [s]	maicena
taatsi [s]	milpilla
tatsi [s]	maizal
tats+ni [s] ts+xi	sacerdote católico
tats+ni [s] ts+xi	cura
tats+ni [s] ts+xi	fraile
tatsiu [s] r+xi	conejo
tatsunáatsi [s]	San José
tatsunáatsi [s]	el mes de Enero
tatsunáatsi [s]	cuando se cambian los oficiales del pueblo
tata [s] ma	padrino
tatatsi [s] ma	tío
tatáata [s]	Jesucristo
tatá [s] te	tendones
tatáame [s]	dios que vive en el lugar donde sacaban pintura blanca para la cara
tatáatsi [s] ma	tío
tatawéeme [s]	luciérnaga
taiwámetsixi [s]	lueciérnagas
tatei kíe [s]	el pueblo de nuestra madre, San Andrés Coahmiata, Jalisco
tatewarí [s]	nuestro abuelo, el dios del fuego
táu [s]	sol

wixárika	español
táutsi [s]	fierro que produce chispa
tautsíxa [s] te	flor amarilla del mes de octubre
tautsíya [s]	chispa
táuxi [s]	pintura facial
tawáari [s] te	huevo
tawáari [s] te	blanquillo
tawarí	otra vez
tawexík+a [s]	nombre del sol
tawí [s] te	pecho
ta+kíkui [s]	pájaro chico de pecho rojizo
taxáta [s] te	tiempo seco
taxárik+ [s] te	estación seca
táraw+kári [s] xi	abeja de colmena real
tayéu [s]	nombre del sol
té [s]	granizo
téka [s] te	piedra vidriosa
téka [s] te	piedra volcánica
téka [s] te	volcán
téeka [s] te	pozo de tatemar
tekata [s]	arroyo con cuevas sagradas cerca de Santa Catarína, Jalisco
'aitsárie [s]	arroyo con cuevas sagradas cerca de Santa Catarína, Jalisco
tekí [s] ri	techalote
tek+ [s] ri	ardilla gris
tek+a [s]	salto
tek+a [s]	cumbre
tekímuxú [s] ri	calabacita
tek+xi [s] te	jarro
matsú [s] te	vaso
tek+xi [s] te	taza
tei [s]	tía
tetéima [s]	tías
téik+a [s]	cumbre
téik+ [s] te	cima de una cuesta
teik+máana [s] te	cima de una cuesta
teiwári [s] xi	vecino
teiwári [s] xi	mestizo
teiwáari [s] ma	madrastra
teiwáari [s] xi	piedra que representa la visita de un pariente, la cual reclama sacrificio para devolver la salud al visitado
tetsú [s]	tamal de frijol y sal para la fiesta del esquite en enero
temáik+ [s] xi	muchacho
temáik+ [s] ts+xi	muchacho
temawíerika [s]	alegría
temári [s]	muchachos
temú [s]	sapo
temú [s]	rana
tem+xiki [s]	polvo obtenido por descascaramiento
téta [s] te	boca
tenúxa [s]	disco de piedra con cavidad en el centro
tepári [s]	piedra que tapa el pozo ceremonial
tépu [s] te	tambor
tep+ [s] tsi	pulga
tep+a [s] te	hierro
tep+tea [s] te	herramienta
tep+a [s] te	fierro
tep+a [s] te	metal
téep+ríki [s]	piedritas
tep+ríkipa [s]	lugar sagrado cerca de San Andrés Coahmiata, Jalisco
ter+ [s] te	cueva
ter+wárika [s]	nombre
ter+wárika [s]	estudio
téetsu [s] te	tamal
téni [s] te	labio
téeta [s] te	boca
teté [s] xi	piedra
teté [s] xi	ídolo
teté [s] xite	piedra
teté [s] xite	ídolo
teukari [s] ma	nieto
teukari [s] ma	abuelo
teukari tsiya [s] ma	el que da nombre
téupa [s]	lugar donde hay mucha pidera

wixárika	español
teukíya [s]	cementerio
teukíya [s]	panteón
teukíya [s]	camposanto
teuxárí [s]	dios que se manifiesta en forma de venado
teuxári [s]	florecita roja
téuri [s] te	muslo
teuríxa [s]	bisapol
teuríxa [s]	bisapol
meripa+t+ [s]	antepasado
tewá [s] má	animal domestico
téewapai	lejos
tewarí [s] ma	nieto
tewarí [s] ma	abuelo del hombre
tewáatsi [s]	payaso ceremonial
téwi [s]	persona
téwi [s]	gente
téwi [s]	indígena
téwi [s]	indio
téwi [s]	autóctono
te+téri [s]	personas
te+téri [s]	indígenas
te+téri [s]	indios
te+téri [s]	autóctonos
xaip+r+ka [s]	papelillo
tupiriya xetá [s]	árbol de corteza roja
xeipiríkari [s]	papelillo
xeipiríkari [s]	árbol de corteza roja
kewiyexa [s] te	huracán
téuka [s] te	remolino grande
texupáme [s] te	resortera
texúuri [s] ma	bisabuelo
texúuri [s] ma	antepasado
teyeupáni [s] te	templo
teyeupáni [s] te	iglesia
tikuékuewáme [s] te	limosnero
tikuikáme [s]	enfermo
tekuiku+kate [s]	enfermos
tiet+	tal vez
tiet+	seguro
tiet+	a poco
tiwawáame [s] te	cobrador
tíxa+	no
tíxa+	nada
tixá+tí	algo

wixárika	español
t+x+r+wáme [s] te	esófago
tiyumiekame [s]	asesino
tiyumiekame [s]	matador
teyuku+káte [s]	asesinos
teyuku+káte [s]	matadores
ti'etsáame [s]	cocinera
ti'etsáame [s]	olla
ti'etsáame [s]	sembrador
tuapúrie [s]	Santa Catarina, Jalisco
tuapúxa+ [s] te	tobillo
tuaxá [s]	roble
xiu [s]	encino roble
túuka [s]	medio día
tuká [s] tsi	araña
tukárik+ [s]	día
tukárí [s]	vida
tukáati [s]	almacenamiento de maíz
túuki [s] xi	camarón chico
t+k+ [s] xi	camarón chico
tuí [s]	arbusto que da pintura amarilla
túixu [s] ri	puerco
tumá [s]	raíz silvestre
tsakuéni [s]	raíz silvestre
tumini [s]	moneda
tumini [s]	dinero
tumuánari [s]	polvo de la tierra
tumuáni [s]	polvo de la tierra
tumía [s]	polvo de la tierra
túnik+ri [s]	círculo
túnikirí [s]	embellecimiento artístico
tunú [s] te	rodilla
túnuáme [s] te	cantador
túnuwamé [s] te	el planeta Vénus
tupí [s] te	arco de flecha
tupí [s] te	arco de violín
tupiri [s] ts+xi	topil
tupiri [s] ts+xi	policía
tupiríya [s] te	hierba
túki [s] te	templo huichol
túru kuaxí [s]	dios que vive en la piedra
turirí [s] xi	codorniz

wixárika	español
tur+wáme [s] te	pato silvestre
túru [s] ts+xi	toro
turú [s] te	articulación
turúxa+ [s] te	articulación
turú [s] te	tobillo
turúxa+ [s] te	tobillo
turú [s] ma	antepasado más retirado
turúxai [s] ma	antepasado más retirado
tutsí [s] ma	padre del abuelo
tutsí [s] ma	madre del abuelo
tutsí [s] ma	padre de la abuela
tutsí [s] ma	madre del a abuela
tutsí [s] ma	hijo del nieto
tutsí [s] ma	hija del nieto
tutsí [s] ma	hijo de la nieta
tutsí [s] ma	hija de la nieta
tutsí [s] ma	tío del abuelo
tutsí [s] ma	tía del abuelo
tutsí [s] ma	tío de la abuela
tutsí [s] ma	tía de la abuela
tutsí [s] ma	hijo del nieto de un hermano
tutsí [s] ma	hija del nieto de un hermano
tutsí [s] ma	hijo de la nieta de un hermano
tutsí [s] ma	hija de la nieta de un hermano
tutsí [s] ma	hijo del nieto de una hermana
tutsí [s] ma	hija del nieto de una hermana
tutsí [s] ma	hijo de la nieta de una hermana
tutsí [s] ma	hija de la nieta de una hermana
tutsí [s] ma	clase de dioses que se maniefiestan en forma de venado
tutsí [s] xi	figuras de masa para uso ceremonial
tutsiwínu [s]	sotol
tutána [s]	tuétano
tutána [s]	médula de hueso

wixárika	español
tutú [s] ri	flor
tutú [s] ma	deidad
tutú [s] xi	costumbre antigua
tutuwí [s]	perico amarillo costeño
tutuwí [s]	dios que se manifiesta en forma de venada
tuwáaxa [s] te	paño de hombre
tuwáaxa [s] te	trapo
tuwár+ [s]	insecto del agua
tuxéeri [s] ts+xi	novio
tuxéeri [s] ts+xi	novia
tuxéeri [s] ts+xi	ilícito
t+ [s]	tizón
t+ [s]	brazas
t+káakame [s]	dios de la muerte
t+káari [s] te	noche
t+k+ [s] xi	camarón chico
túuk+ [s] xi	camarón chico
t++pína [s]	chuparrosas
t+raméka [s]	cerro sagrado cerca de Huajimic, Nayarit
t+ráanari [s]	trueno
t++rí [s] xi	niños
t++rí [s] xi	hijos
t+ríkukúuyame [s]	hierba que hace salir pedazos de un hueso quebrado a los cinco días
t+r+kíta [s]	dios que enferma a los niños
t+riyáma [s] ma	familia
t+r+karíya [s]	fuerza
t+reku+yu [s]	oficial encargado de la imagen de Jesucristo
t++wáinu [s] r+xi	niño que suena la sonaja en la fiesta del tambor
t+we [s] xi	tigre
hukurí [s] xi	gavilán de cola blanca
t+xáari [s]	carbón
t+ [s]	brazas

wixárika	español
t+xí [s]	masa
wakána [s] ri	gallina
turukí [s] ri	El Carro
wakáxi [s]	vaca
wakáitsixi [s]	vacas
wái [s] te	carne
waikári [s] te	juego
waikári [s] te	juguete
waikári [s] te	objeto ceremonial
waikamé [s]	jugador
wáinu [s]	calandria
wáinu [s]	pajarito amarillo
máana	allí
kwap+ [s] te	pezuña de calabaza
uwakí [s]	nanchi
wárie [s] te	espalda
waríe	pasado mañana
waríena [s]	destrás de él
waritsi [s] tsixi	enjambre
waritsi [s] tsixi	especie de avispa
warúutsi [s] ma	madre
waríka+ [s] ma	suegra del hombre
wár+i [s] te	columpio
wár+i [s] te	cuna de columpio
wátsie [s]	nombre de un dios
watá	allí
inet+arika [s] te	espinilla
watúuxa [s]	mariposa que aparece al final de la estación de lluvias
kuateří [s] xi	gemelos
kuateří [s] xi	cuates
wawatsári [s]	dios que se manifiesta en forma de venado
wawá+ri [s]	bola de músculo en el cuerpo
wáawe [s]	chual
wawéeme [s] te	abejón
wa+ká	mucho
wa+káwa	mucho
wá+t+a [s]	San Sebastian, Jalisco
wa+riyárika [s]	a fuerza
wáxa [s]	milpa
wátsiya [s]	coamil

wixárika	español
waxíewe [s]	dios que vive en una priedra en el mar cerca de San Blas, Nayarit
hayewáxi [s] te	guayaba
wa'at+	parece
wa'at+	quizás
wéiya [s]	Abril
weříka [s] xi	águila
weříka [s] xi	avispa grande
weříka [s] xi	zángano
weříka [s] xi	cantador que participa en la fiesta del tambor como ayudante
weříka '+imári [s]	dios que habita en el cielo
we+ra'+ká [s]	planta de la que se hacen flechas
amupa,amunena [s]	el más grande
mat+ří [s]	el principal
wíya [s]	año
wíexu [s] ri	malacoa
winíyáari [s] te	lazo para coger venados
wipí [s] te	red
wikí [s] xi	pájaro
w+rikúta [s]	tierra del peyote
w+rikúta [s]	el oriente
w+rikúta [s]	donde termina la tierra
wir+k+ [s] xi	zopilote
witsexá [s]	sierrilla
witsaxá [s]	guía espinosa
witsée [s] ri	pájaro chico que mata otros pájaros
w+tséx+ka [s]	dios que vive cerca de Zacatecas
tsip+ ane	bonito
wits+ánari [s]	tela pegajosa
wíta [s] te	estambre de lana
w+tári [s]	tiempo de aguas
w+tári [s]	tormenta
w+yéri [s]	lluvia
wíte,witwya [s]	hachazo
wituríxi [s]	dios que vive en un peñasco

wixárika	español
w+wíeri [s]	aza
w+wíeri [s]	colgadera
w+wíeri [s]	cordón para colgar
w+xáarika [s]	Huichol
w+xáritari [s]	Huicholes
wíxi [s] ma	sobrino del esposo
wíxi [s] ma	sobrina del esposo
wíxi [s] ma	esposa de un tío
wíxi [s] ma	tía política
wíxi [s] ma	concuña
wiyá [s]	aceite
wiyá [s]	manteca
xakwitsári [s]	nixtamal
xak+	plato de barro
xak+r+te [s]	platos de barro
xak+pári [s] te	tepalcate
xaim+ari [s]	menudo
xaim+ari [s]	estómago de animal
xaim+ari [s]	talega ancha
xainiú [s]	víbora ponzoñosa
xaipir+kari [s]	papelillo
xaipir+kari [s]	árbol de corteza roja
téwi xetá [s]	papelillo
téwi xetá [s]	árbol de corteza roja
xáip+ [s] tsi	mosca
xamá [s] te	hoja de elote
m+axá [s] te	hoja de milpa
xápa [s] te	papel
xápa [s] te	chalate
xápa [s] te	zalate
xapawiyemeta [s]	dios que vive en un lago al sur
xapí [s]	vagina
xapí [s]	órgano femenino
xapúni [s] te	jabón
xakí [s]	esquite
xakí [s]	maíz tostado
xáari [s] te	olla
xáatsi [s]	basura
xáatsi [s]	desperdicio
xatá [s] te	jícama
xat+ [s] te	comal
xat+ [s] te	rótula
xat+ [s] te	especie de gavilán
xat+pári [s] te	pedazo de comal quebrado

wixárika	español
xat+pári [s] te	hojalata
xamá[s] ri	hoja
xáawe,karimutsi, [s]	el arbol y su fruto pochote
xaweréru [s] ts+xi	músico
xawéri [s]	violín
xawéri [s]	música
xawerúxi [s] te	calzones
xa+tá [s]	sólo
xatsíka [s]	sólo
xáaye [s] tsi	víbora de cascabel
tumatí [s] te	jitomate
tumati	tomate colorado
tats+h+xí [s] te	
xéek+i [s] tsi	jején
xeik+a	solamente
xeik+a	nada más
xeimíeme [s]	una vez
xeitapáari [s]	una parte
xeitéwiyári	veinte
xemúutsi [s] te	ano
xepái [s] te	hierva de venado
xerái [s] ma	esposo del nieto
xerái [s] ma	esposa del nieto
xerái [s] ma	esposo de la nieta
xerái [s] ma	esposa de la nieta
xerái [s] ma	abuelo del esposo
xerái [s] ma	abuela del esposo
xerái [s] ma	abuelo de la esposa
xerái [s] ma	abuela de la esposa
xerái [s] ma	talega grande de lana
xéri [s]	frío
xetakuakuáxi [s]	granadillo
xetat+kuakuáaxi [s]	granadillo
xetáríka [s]	pintura facial
xewá [s] te	hoja de calabaza
xewí [k]	uno
xewít+ [k]	uno
xewít+ [k]	unos
h+pát+ [k]	uno
h+pát+ [k]	unos
xikúri [s] te	jolote
xikúri [s] te	ques quémil
x+k+a [s]	pesuña trasera de venado

wixárika	español
x+k+a [s]	lizo del telar
xik+ri [s] te	espejo
xiekári [s]	arena
xíete [s] xi	miel
xíete [s] xi	abeja de colmena real
wiurí [s]	especie de gavilán
x+méri	temprano
x+méri	la mañana
atakwai [s] xi	lagartija que vive en los árboles
xikiúnipa [s]	lugar sagrado cerca de San Andrés Coahmiata, Jalisco
x+ríki [s]	casa ceremonial chica
xitá [s] xi	jilote
xitá [s] xi	brote de elote
xiurí [s] tsi	ajolote
xíxi [s]	orina
xukúuri [s] te	jicara ceremonial
xumáatsi [s]	especie de avispa
xupuréeru [s] te	sombrero
xuráwe [s] ts+xi	estrella
xur+ya [s]	sangre
xútsi [s] te	calabaza
xuiyá [s]	bordado
xutúri [s]	flor
xutúri [s]	flor de papel
xuxúi [s]	grillo
xuxúi [s]	dios que se manifiesta en forma de venado
xuyá [s] te	espina
xuyári [s]	palo espinoso
xur+ya [s]	sangre
x+ka	si
x+mí [s] te	anona
x+mí [s] te	chirimoya
x+nai [s] tsi	liendres
t+niriya [s] tsi	caspa
x++nári [s]	adobe
x+náríta [s]	dios que vive en un peñasco
x+natsáta [s]	lugar angosto entre piedras partidas

wixárika	español
x+ri [s]	calor
x++rí [s] xi	huizache
x++rí [s] xi	palo espinoso
x+rikíya [s]	zigzag
x+r+ka [s]	panal
x+ríka [s]	cera
x+té [s] te	uña
x+té [s] te	garra
x+témútsi [s] te	ombligo
x+téetema [s]	figuras de pinole hechas con panocha para la fiesta del Carnavalo de la Virgen de Guadalupe
xíye [s] tsi	armadillo
yá [s]	tabaco
yá	o
wa+ka mex+a	varias veces
wa+kamex+a	repetidamente
yák+	repetición
yatéwa [s]	periente lejano
yáawi [s] xi	coyote
yéekuai [s] te	bule para tabaco sagrado
yakíri [s]	cráneo
yéimukuáari [s]	zacate amarillo para prender lumbre
yeiyári [s] te	costumbre
yeiyári [s] te	tradición
yemúri [s] te	cerro
yéeri [s] te	camote
yéeri [s] te	guacamole del monte
yetána [s]	arriba de él
yeturíxa [s] ma	pariente que se llamaba cué, del cual murió el esposo
yeturíxa [s] ma	pariente que se llamaba cué, del cual murió la esposa
yéuka [s] te	aguacate
yéuparetá	palabra con que se terminan cuentos
yéuxu [s] ri	tlacuache
ye'+ [s] te	cántaro
ye'+ [s] r+te	cántaro

wixárika	español
yúari [s] xi	guacamaya
yuaríya [s]	ruido
yuawíme [s]	maíz azul
yuheyéme [s]	siempre
yuheyéme [s]	totalmente
yuheyéme [s]	de a tiro
yuimakuári [s]	fiesta del tambor
yuimakuári [s]	fiesta de calabacitas
yúri	verdad
yúri	la verdad
yurienáka [s]	dios de la tierra mojada
yuríepa [s] te	estómago
yuríepa [s] te	barriga
yuríepa [s] te	vientre
y+k+	distinto
y+k+	diferente
y+k+	otro
y++ná [s] te	tuna
y++rári [s]	retoño
tsikuraté [s]	chocolate ceremonial
y++ríya [s]	oscuridad
'apa	grande
'apat+	de cierto tamaño
'anet+	grande
'wa+kawá	de cierto tamaño
'akúxi	todavía
'tewiyariká [s] te	peña
'ai [s] te	peñasco
'aikútsi [s] te	jícara para ofrendas de tejuino
'aikútsi [s] te	tecomate
'áik+ [s]	el año pasado
'áik+ [s] xi	asquel
'áik+ [s] xi	hormiguita
'áina [s] ri	cangrejo
'aitsári [s]	arroyo cerca de Santa Catarina, Jalisco, donde hay cuevas sagradas
'tekáta [s]	arroyo cerca de Santa Catarina, Jalisco, donde hay cuevas sagradas

wixárika	español
'aitsárite [s]	dioses que viven en 'Aitsárie
'aitsíka [s]	mandamiento
'aitaráma [s]	víbora parecida a la coralilla
'áite+yá [s]	eco
'áix+a	bueno
'áix+a	limpio
'áana	entonces
'ánaké	entonces
'aná [s] te	ala
'aná [s] te	pluma de flecha
'anirú [s] te	anillo
'tsiwerí [s]	árbol silvestre
'áki [s] te	arroyo
'áki [s] te	costilla
'arí	ya
'ariké	después
'áaru [s] xi	guajolote
'áaru [s] xi	pavo
'átsi [s] xi	murciélago
'kukaratsa [s] xi	cucaracha
'atákuai [s] xi	lagartija
'atákuai [s] xi	iguana chica
'ataháika [k]	ocho
'atahúuta [k]	siete
'atanáuka [k]	nueve
'atári [s] te	testículos
'ataxewí [k]	seis
'até [s] tsi	piojo
'áatu	anteayer
'áatu	antier
'aukuér+ka [s]	dios que vive en un cerro de San Sebstián, Jalisco
'ayekwai [s] te	quijada
'a+rí [s] te	mejilla
'hurá	cerca
'á+ri [s] te	cachete
'á+ri [s] te	pómulo
'a+x+wi [k]	cinco
'áawa [s]	tumor del hocico en los caballos
'awá [s] te	cuerno
'a+riká [s]	saliva
'á+te [s] tsi	hormiga colorada que pica

wixárika	español
’á+xi [s]	vapor
’áxa	malo
’áxa	sucio
’a xeik+a	juntos
’a xeik+a	en uno
’áxi [s] ma	hermano menor de la mujer
’áxi [s] ma	hermana menor del hombre
’ax+kái	despacio
’axuxí [s]	ajo
’ayé [s] tsi	tortuga
’ayékuai [s] xi	quijada
’ayékuai [s] xi	barbilla
’ayeimána	postpasado mañana
’ayepári [s] te	tepetate
’ayepári [s] pa	tepetatera
’éekáa [s]	viento
’éekáa [s]	aire
’ék+	tú
’ék+	usted
’éna	aquí
’éri	sí
’éri [p	como no
’etsá [s]	grano
’etsá [s]	viruela
tsimpe	chico
étsi	feo
’et+riya [s] te	sombra
’iká [s] te	carga
’íkwai [s]	comida
’tikwaiwame [s]	fruta
’ík+	éste
’ík+	ésta
’imá [s]	taco
’imat+réeme [s]	el siguiente
’imát+riéka [s]	el último
’+múmui [s] te	escalera
’im+ari [s] te	semilla
’im+kui [s]	ajolote
’ín+a [s] te	bastimento
’in+arí [s]	muestra
’ín+ari [s]	medida
’ípa+	así
’+pá+ [s]	víbora mitológica
’+níya [s]	tumba

wixárika	español
’xawariyá [s]	hueco
’ipítsa [s]	cetro de zacatón
’ipurí [s] te	bola de estambre
’ipurí [s] te	hilaza
’+ki [s]	granero para maíz
’ikitsára+ [s] te	horcón
’ikitsára+ [s] te	horqueta
’ir+karíya [s]	tosferina
’ir+karíya [s]	pertusis
’itsáari [s] te	caldo
’itsáari [s] te	comida
’ítsi [s] te	petate
’itsik+na [s] te	esquina
’íts++ [s]	bastón ceremonial
’íts++kame [s]	señal de autoridad
’its+ári [s]	palo para batir nixtamal
’its++káme [s]	juez
’its++káme [s]	gobernador
’its++káte [s]	jueces
’its++káte [s]	gobernadores
’+tári [s]	tejido ceremonial
’+tári [s]	sudadera
’itárumari [s] te	majahua
’itárumari [s] te	fibra
’ité+ri [s] te	arbusto
’ité+ri [s] te	cosa plantada
’itiwáame [s] te	escoba
’itúa [s] te	nido
hítua [s] te	nido
’itúupari [s]	puerta
’itúupari [s]	tapaderade la puerta
’itúupiri [s] te	tapón
’itúupiri [s] te	tapadera
’ít+ [s] te	cuchara
’it+wáame [s]	dedo
’iwá [s] ma	hermano
’iwá [s] ma	hermana
’iwá [s] ma	primo hermano
’iwá [s] ma	prima hermana
’iwá [s] mar+xi	hermano
’iwá [s] mar+xi	hermana
’iwá [s] mar+xi	primo hermano
’iwá [s] mar+xi	prima hermana
’iwarú [s] ma	cuñada de la mujer

wixárika	español
'íwi [s] te	falda
'íwi [s] te	enaguas
'iwieyáme [s] te	golondrina
'iwipáme [s] te	alfiler
'iwipáme [s] te	aguja
'ixáuri [s]	pluma que se pone en el sombrero
'ixuríki [s] te	ropa
'ix+anáka [s]	Vía Láctea
'iyá	aquél
'iyá	ése
'iiyá [s] te	pulmón
'itukariyá [s] te	bofe
'+yá [s] te	resuello
'iyáari [s] te	corazón
'iyáari [s] te	alma
'uakáxa [s] te	rastrojo
'uakáxa [s] te	canyecote
'uká [s] ri	mujer
'ukarái [s]	vieja
'ukarái [s]	esposa
'ukaráatsi [s]	viejita
'ukaráawetsixi [s]	viejitas
'úukai [s]	lágrimas
'úha [s] te	caña
'uháye [s] te	medicina
'uháye [s] te	veneno
'uháye [s] te	remedio
tiyu'uhayemáawáame [s]	curandero
tiyu'uhayemáawáame [s]	médico
tiyu'uhayemáawáame [s]	doctor
'umé [s] te	hueso
'úna [s]	sal
'upára+ [s] te	sosopastle
'upára+ [s] te	machete del telar
'ukí [s] tsi	hombre
'ukirái [s]	viejo
'ukirái [s]	marido
'ukirí [s]	gallo
'ukiráatsi [s]	viejito
'ukiráawets+xi	viejitos
'ukiyáari [s] ma	padre
'ukiyáari [s] ma	jefe

wixárika	español
'ukiyáari [s] ma	oficial
'ukiyáari [s] ma	partrón
'ukiyáriwáari [s]	padrastro
'utsí [s]	ocote
'utsika [s] ri	chapulín
'utsík+i [s]	huso de hilar
'utsúyena [s]	pipa ceremonial
'utá [s] te	cama
'utamána [s]	su izquierda
'ututáwi [s]	dios que se manifiesta en forma de venado
'ut+anáka [s]	dios que vive en 'Aitsárie, que hace reveinta los tímpanos
'ut+mána [s]	atrás de él
'ut+arika [s]	escritura
'umá	en otra parte
'úwa	aquí
'uwakí [s] te	nanchi
'uwéeni [s]	sillón de otate
'uxa'á	mañana
'uximayatsíka [s]	trabajo
'uyúuri [s] te	cebolla
'xukurí [s]	jícara
'kuxaurí [s]	bule no cortado
'+ká [s] te	pierna
'+ká [s] te	pie
'+kári [s] te	cobija
'+kua [s] te	chicle
'+kua [s] te	copal
'+kua [s] te	resina para el violín
'+kuáari [s]	quemazón
'íkwi [s]xi	lagartija
'++kíi [s] te	palmera
'++k+i [s] te	gancho para cortal palma
'+imári [s] xi	muchacha
"+pá [s] te	zorrillo
'++pári [s] te	banco
'++pári [s] te	silla
'++kítsika [s] te	patrón
'++kítsika [s] te	dibujo
'++kítsika [s] te	muestra
'++kít+ariká [s] te	enseñanza

wixárika	**español**
'++kit+arika [s]	patrón
'++kit+arika [s]	dibujo
'++kit+arika [s]	muestra
'++kit+arika [s]	enseñanza
'+kí [s]	mazorca de semilla que se guarda para la fiesta de la siembra
'+ra [s] xi	correcaminos
'+ra [s] xi	pájaro flojo
'+rawí [s] xi	correcaminos
'+rawí [s] xi	pájaro flojo
'+r+ [s] te	flecha
'+tsa [s]	cubierta para flechas
'+tsá [s]	brasil
'++tsi [s]	monte
'++tsi [s]	maleza
'+t+rai [s] xi	zancudo
'++wári [s] te	baño
'++wí [s]	nieve
'+xá [s] te	zacate
'+xá [s] te	pasto
'+xáatsi [s] te	cuento
'+ya [s]	esposa
'+itáma [s]	esposas
'iyáari [s] ma	ahijado

Apéndice C

Corpus apareado

A continuación se presenta el conjunto de frases apareadas en wixárika y español que fueron utilizadas para entrenar el traductor. Las frases han sido retomadas únicamente con fines académicos de la obra "Huichol de San Andrés Cohamiata, Jalisco" de Gómez, Paula Gómez (1999). La versión presentada no cuenta con segmentación morfológica, pero puede ser encontrada segmentada en la obra original, al igual que con anotaciones de cada morfema.

Tabla C.1: Corpus paralelo wixárika - español

wixárika	**español**
'ik+ ki 'ep+pa	esta casa no es grande
'ik+ ki 'ep+kapa	esta casa no es grande
'ik+ ki tsip+pe	esta casa es chica
'ik+ ki tsip+kape	esta casa no es chica
'ik+ ki p+hekwa	esta casa es nueva
'ik+ ki p+kahekwa	esta casa no es nueva
'ik+ ki p+'ukiratsi	esta casa es vieja
'ik+ ki p+ka'ukiratsi	esta casa no es vieja
'ik+ ki p+tuxa	esta casa es blanca
'ik+ ki p+katuxa	esta casa no es blanca
m+k+ ki kehe'ane	¿como es esa casa?
m+k+ ki kehepa	¿que tan grande es esa casa?
ki m+k+ tihekwa	¿es nueva esa casa?
h+ p+hekua	si es nueva
hawaik+ p+kahekwa	No,no es nueva
h+ hawaik+	si no
ne 'aneputeewi	yo soy alto
'ek+ 'apeputeewi	tu eres alto
m+k+ 'aputeewi	el eres alto
m+k+ 'aputeewi	ella ere alta
tame 'ateput+t+	nosotros somos altos
xeme 'axeput+t+	ustedes son altos
m+k+ 'ameput+t+	ellos son altos

wixárika	**español**
ne tsinepawe	yo soy chaparro
'ek+ tsipepawe	tu eres chaparro
m+k+ tsipawe	el es chaparro
m+k+ tsipawe	ella es chaparra
tame tsitepa'u	nosotros somos chaparros
xeme tsixepa'u	ustedes son chaparros
m+k+ tsimepa'u	ellos son chaparros
m+k+ tsimepa'u	ellas son chaparras
ne nep+waiya	yo soy gordo
'ek+ pep+waiya	tu eres gordo
m+k+ p+waiya	el es gordo
m+k+ p+waiya	ella es gorda
tame tep+waiyat+ka	nosotros somos gordos
xeme xep+waiyat+ka	ustedes son gordos
m+k+ mep+waiyat+ka	ellos son gordos
m+k+ mep+waiyat+ka	ellas son gordas
ne nep+waki	yo soy flaco
'ek+ pep+waki	tu eres flaco
m+k+ p+waki	el es flaco
m+k+ p+waki	ella es flaca
tame tep+wawaki	nosotros somos flacos
xeme xep+wawaki	ustedes son flacos
m+k+ mep+wawaki	ellos son flacos
m+k+ mep+wawaki	ellas son flacas
neki 'ep+pa	mi casa es grande
'aki 'ep+pa	tu casa es grande
m+k+ kiya 'ep+pa	la casa de él es grande
taki 'ep+pa	nuestra casa es grande
xeki 'ep+pa	la casa de ustedes es grande
waki 'ep+pa	la casa de ellos es grande
kukuri p+xeta	el chile es rojo
kukuriri p+xeta	los chiles ya estan rojos
kukuri p+kaxeta 'akuxi	los chiles todavia no estan rojos
m+k+ k+ye 'ep+pa	ese árbol es grande
m+k+ k+ye ep+tapare	ese árbol va aser grande
m+k+ k+ye 'ep+pakairi	ese árbol ya era grande
m+k+ k+ye 'eputaparix+	ese árbol se volvió grande
kukuri p+tixetare	los chiles se van a volver rojos
'ik+ xupureru 'axupureru hepa+ p+'ane	este sombrero es igual al tuyo
'ik+ xupureru 'axupureru hepa+ p+ka'ane	este sombrero no es igual al tuyo
'ik+ xupurerute y+k+ p+'anene	estos sombreros son diferentes
m+k+ 'iwi ke'ane	¿de qué color es esa falda?
m+k+ 'iwi y+wit+ p+tuxa	esa falda es negra y blanca
'akawayu reuy+wi nutsu peutuxa	¿tu caballo es negro o blanco?
nekawayu p+kaheuy+wi p+kaheutuxa peuwayu p+ta	mi caballo no es negro ni blanco es, bayo

wixárika	**español**
'ikwai keha'ane	¿cómo esta la comida?
'ikwai ke'ane	¿cómo esta la comida?
'ikwai	comida
'ikwai p+x+ka	la comida está caliente
ikwai p+ha+t+	la comida está fría
ha kwinie p+tiha+t+	el agua está muy fría
ha wa+kawa p+ha+t+	el agua esta bastante fría
ha tsip+katiha+t+	el agua esta demaciado fria
m+k+ xari 'axa p+'ane	esa olla esta sucia
m+k+ xari 'axa p+waku'ane	esa olla esta sucia
m+k+ xari 'axa p+ka'ane	esa olla no esta sucia
m+k+ xari 'axa p+kawaku'ane	esa olla no esta sucia
m+k+ xari 'aix+ p+'ane	esa olla esta limpia
m+k+ xari 'aix+ p+waku'ane	esa olla esta limpia
m+k+ xari 'aix+ p+ka'ane	esa olla no esta limpia
m+k+ xari 'aix+ p+kawaku'ane	esa olla no está limpia
kiekaritsie pehura 'ena	el pueblo está cerca de aquí
kiekaritsie petewa 'ena	el pueblo está lejos de aquí
kiekari petewa	¿está lejos del pueblo?
wani hik+ tateikie peka	juan está ahora en san andrés
hakewa peka hik+ wani	¿donde está juan ahora?
hik+ tateikie reka	¿ahora está en san andrés?
tateikie reka m+k+	en san andrés es donde está?
wani takai tateikie peyeikakai	ayer juan estuvo en san andrés
hakewa wani takai peyeikakai	¿dónde estuvo juan ayer?
hakewa 'apapa	¿donde esta tu papá?
yukie reka 'apapa	¿tu papá esta en la casa?
yukie 'apapa kareka	¿tu papá no esta en la casa?
takie nepapa puka	mi papá está en la casa
nepapa takie p+ka'uka	mi papa no esta en la casa
nepapa nekie ya p+tiuka	mi papá siempre está en la casa
nepapa waxata p+kayeika	mi papá está en la milpa
nepapa mercado payeka	mi papá está en el mercado
nepapa hat+a peyeka	mi papá esta en el río
nepapa manuwerits+a peka	mi papá está con don manuel
kem+'ane hets+a 'apapa peka	¿con quién está tu papá?
'apapa muwa rayeka	¿está tu papá?
hakewa xari	¿dónde está la olla?
hakewa kutsira	¿dónde está el machete?
xari kwiepa puka	la olla está en el suelo
kutsira kwiepa puka	el machete esta en el suelo
xari huta putika	la olla está en ese rincon
kutsira huta putika	el machete está en ese rincon
xari xat+ 'aurie puka	la olla está junto al comal
hakewari ts+k+	¿dónde está el perro?
ki 'aurie puwe	está fuera de la casa

wixárika	**español**
kita payewe	está dentro de la casa
hakewa teyupani peh+k+	¿dónde está la iglesia?
hakewa pemerkato	¿dónde está el mercado?
hakewa 'aki pewe	¿dónde está tu casa?
hakewa pe'akie	¿dónde está tu casa?
teyupani 'iya ki h+xie puwe	la iglesia está enfrente de aquella casa
teyupani 'uma pai 'etsiwa pewe	la iglesia esta mas adelante
hakewa kuraru pema	¿donde está el corral?
ki warie kuraru puma	el corral está atrás de la casa
ki h+xie kuraru puma	el corral está adelante de la casa
ki aurie kuraru puma	el corral está junto a la casa
kemari baultsie patika	la ropa está en el baul
mume xarita p+yema	los frijoles están en la olla
wiki k+yetsie puyeka	el pájaro está en ese árbol
xaip+ tekitsie paka	la mosca está en la pared
ts+ik+ri tai 'aurie mepatet+ka	los perros están alrededor de la lumbre
ts+ik+ri kepa+ pep+warexeiya	¿cuántos perros tienes?
ts+ik+ri kepa+ pep+watewa	¿cuántos perros tienes?
ts+k+ xeime nepexeiya	tengo un perro
ts+k+ xeime nep+tewa	tengo un perro
ts+k+ xeime 'ek+ pemexeiya	tienes un perro
ts+k+ xeime pem+tewa	tienes un perro
m+k+ xeime ts+k+ m+tewa	el tiene un perro
m+k+ xeime ts+k+ pexeiya	el tiene un perro
tame xeime ts+k+ tepexeiya	tenemos un perro
tame xeime ts+k+ tep+tewa	tenemos un perro
xeme xeime ts+k+ xepexeiya	ustedes tienen un perro
xeme xeme ts+k+ xep+tewa	ustedes tienen un perro
m+k+ xeime ts+k+ mepexeiya	ellos tienen un perro
m+k+ xeime ts+k+ mep+tewa	ellos tienen un perro
ts+k+ri meyuhutame nep+warexeiya	tengo dos perros
ts+k+ri meyuhutame nep+watewa	tengo dos perros
ts+k+ri meyuhutame pep+warexeiya	tienes dos perror
ts+k+ri meyuhutame pep+watewa	tienes dos perror
m+k+ ts+k+ri meyuhutame p+warexeiya	el tiene dos perros
m+k+ ts+k+ri meyuhutame p+watewa	el tiene dos perros
tame ts+k+ri meyuhutame tep+warexeiya	tenemos dos perros
tame ts+k+ri meyuhutame tep+watewa	tenemos dos perros
xeme ts+k+ri meyuhutame xep+warexeiya	ustedes tienen dos perros
xeme ts+k+ri meyuhutame xep+watewa	ustedes tienen dos perros
m+k+ ts+k+ri meyuhutame mep+warexeiya	ellos tienen dos perros
m+k+ ts+k+ri meyuhutame mep+watewa	ellos tienen dos perros
kawayu xeime nepexeiyakai	tenía un caballo
kawayu xeime nep+tewakai	tenía un caballo
yurika xeime kawayu nepexeiyani	el año que viene voy a tener un caballo
yurika xeime kawayu nep+tewani	el año que viene voy a tener un caballo

wixárika	**español**
ts+k+ nep+kahexeiya	no tengo perro
nep+ka'ukats+k+	no tengo perro
nep+ka'uyetumini	no tengo ninguna moneda
ha nepexeiya hariwame	tengo agua para tomar
tita muwa petiyepine	¿qué tienes ahí en el costal?
kem+'ane kutsira pakwee	¿quién tiene el machete en el costal?
ne nepiteka	yo lo tengo
'iku p+k+m+tsie	el maíz tiene gorgojo
ts+ik+ri mepakwaxi	los perros tienen cola
ts+ik+ri metehakwaxi	¿tienen cola los perros?
ts+ik+ri tam+ mekatehakwaxit+ka	¿acaso no tienen cola los perros?
nets+k+ p+kahakwaxi	mi perro no tiene cola
'iya ki 'itupari p+kaheuwie	esa casa no tiene puerta
wani 'etsiwa p+rekak+pa	juan tiene poco pelo
wani wa+kawa p+rekak+pa	juan tiene mucho pelo
kepa+ anene wani k+paya	¿como es el pelo de juan?
wani k+paya p+y+y+wi	el pelo de juan es negro
kutsira pem+netsi'uni+t+a neputeka	tengo el machete que me prestaste
kutsira pem+netsi'umi 'akuxi neputeka	todavía tengo el machete que me regalaste
nekie kutsira nepeteka kanekutsira	en casa tengo un machete que no es mío
'ik+ kamixa 'axa peuku'anene	esta camisa tiene manchas
'ik+ ts+k+ peukutuxa	este perro tiene manchas blancas
'ikwai p+'unama	la comida tiene sal
'ikwai p+ka'unama	la comida no tiene sal
'ikwai pa'utsiwi	la comida esta salada
'itsari p+kukurima	la sopa tiene chile
xari ha puyema	la olla tiene agua
xari ha p+ka'uyema	la olla no tiene agua
hakewa wani kaunari peyeyetsa	¿dónde tiene juan el mecate?
wani kaunari pahurie	juan tiene el mecate en la mano
wani yupurutsata tete puyeyetsa	juan tiene una piedra en el bolsillo
yuhutame nep+wa'iwa	tengo dos hermanos
kemey+pa+me 'a'iwama	¿cuántos hermanos tienes?
'ek+ xapuni perexeiya tuiyari	¿tiene usted jabón para vender?
nep+tsukaxie	tengo gripa
neputatsukaxi	tuve gripa
pereuxerim+k+	¿tienes frio?
h+	si
waik+	no
nemu'u p+netsi'ukukwine	tengo dolor de cabeza
nep+mu'ukwine	tengo dolor de cabeza
nep+nemex+it+a	tengo prisa
nepeuhakam+k+	tengo hambre
nepeuharim+k+	tengo sed
nepeukum+k+	tengo sueño
nepeune'+raxie	tengo flojera

wixárika	**español**
nep+reka'uximayatsika	tengo trabajo
nepeunetewiya	tengo pena
nep+netewiya	soy penoso
xarita mume puyema	en la olla hay frijoles
xarita mume tiuyema	¿hay frijoles en la olla?
xarita mume katiuyema	¿no hay frijoles en la olla?
h+ puyema	sí hay
waik+ p+kauyema	no, no hay
h+ritsie maxatsi mekatehexuawe	¿en el monte no hay venados?
kiepa te+teri mep+xuawe	en la casa hay gente
kiepa te+teri mep+kaxuawe	en la casa no hay gente
mana kiepa te+teri mep+kaxuawekai	en la casa no habia gente
'ena paapa naukat+ pamane	aquí hay cuatro tortilla
'ena paapa naukat+ p+yemane	aquí hay cuatro tortilla
'ena paapa xewit+ pama	aquí hay una tortilla
'ena paapa xewit+ p+yema	aquí hay una tortilla
'ena ha p+xuawe	aquí hay agua
'ena ha p+yema	aquí está el agua
'ena ha p+kamawe	aquí no hay agua
'ena ha p+kaxuawe	aquí no hay agua
'ena ha p+kayema	aquí no está el agua
'uma k+yexi p+xuawe	allá hay árboles
kwiniya kiekaritsie p+tiyu'axiya	una epidemia ataca el pueblo
nekie mexikaru pexuawe	en mi pueblo hay mercado
mana kiekaritsie ki kwinie papat+ pexuawe	en ese pueblo hay casas muy grandes
tsanate wikit+t+ p+titewa	el zanate es un pájaro
m+k+ wiki p+tsanate	ese pájaro es un zanate
m+k+ wiki p+katsanate	ese pájaro no es un zanate
m+k+ wiki tsanate p+titewa	ese pájaro es un zanate
m+k+ wiki tsanate p+katitewa	ese pájaro no es un zanate
tita wikiyari tih+k+ m+k+	¿que pájaro es ese?
tita tih+k+ m+k+ wiki	¿que pájaro es ese?
m+k+ wiki ketitewa	¿que pájaro es ese?
wani p+ti'+kitame	juan es maestro
wani p+kati'+kitame	juan no es maestro
kem+'ane p+ti'+kitame	quien es maestro
kem+ane ti'+kitame p+h+k+	quien es el maestro
wani neniwe p+h+k+	juan es hijo mio
wani p+neniwe	juan es hijo mio
wani tita tih+k+	¿qué es juan?
wani p+tiyu kewaiya	juan es brujo
wani tiyu kewaiyame pat+a	juan se volvió brujo
m+k+ p+ne tsik+iwiti	esa canasta es mía
m+k+ p+'atsik+iwiti	esa canasta es tuya
m+k+ p+tsik+iwitieya	esa canasta es de él
m+k+ p+tatsik+iwiti	esa canasta es nuestra

wixárika	**español**
m+k+ p+xetsik+iwiti	esa canasta es de ustedes
m+k+ p+watsik+iwiti	esa canasta es de ellos
m+k+ p+kanetsik+iwiti	esa canasta no es mia
m+k+ p+ka'atsik+iwiti	esa canasta no es tuya
m+k+ p+katsik+iwitieya	esa canasta no es de él
m+k+ p+katsik+iwitieya	esa canasta no es de ella
m+k+ p+katatsik+iwiti	esa canasta no es nuestra
m+k+ p+kaxetsik+iwiti	esa canasta no es de ustedes
m+k+ p+kawatsik+iwiti	esa canasta no es de ellos
m+k+ netsik+iwiti p+h+k+	mi canasta es esa
netsik+iwiti m+k+ p+kah+k+	mi canasta no es esa
kem+ane m+k+ p+ratsik+iwiti	¿de quién es esa canasta?
m+k+ tiatsik+iwiti	¿es tuya esa canasta?
m+k+ atsik+iwiti tih+k+	¿esta canasta es la tuya?
'ik+ uye p+h+k+	es el camino
wani tateikie p+kiekame	juan es de san andrés
tateikie	san andrés
hakewa wani pekiekame	¿de dónde es juan?
hakewa pepekiekame	¿de dónde eres?
kem+'ane tateikie p+kiekame	¿quién es de san andrés?
wani 'aix+ p+tiuka'iyari	juan es un hombre bueno
wani tupiri payani	juan va a ser topil
tita rayaniwani	¿qué va a ser juan?
wani aik+ tupiri pat+a	juan fue topil el año pasado
kepa+ka tupiri pat+a wani	¿cuándo fue topil juan?
tita rat+a wani aik+	¿qué fue juan el año pasado?
wani p+uki	juan es hombre
maria p+uka	maria es mujer
ne nep+uka	yo soy mijer
ne nep+uki	yo soy hombre
tita tiitsari'yari	¿qué es esta comida?
x+ye p+waiyari	es carne de armadillo
ne yak+ nep+nemate	yo soy el más joven
ne yak+ nep+kanemaate	yo no soy el mas joven
maka wani p+h+k+	el que está sentado es juan
wani p+ta maka p+h+k+	juan es el que está sentado
ekitsata cabecera municipal p+h+k+	ezquitic es cabecera municipal
ed+wikes ukari mep+teu ter+war+wa	eduviges es nombre de mujer
uki manuyet+a nemimate p+y+ane	el hombre que salió es al que conozco
manuyet+a nemimate p+y+ane	el que salio es al que conozco
tita tih+k+	¿qué es eso?
p+ts+k+	es un perro
neniwe yurika 'ep+yutamani	el año que viene mi hijo ya va a ser hombre
wani maria hepa+ p+raka'erie	juan se parece a maria
wana yumama hepa+ p+raka'erie	juana se parece a su madre
'ek+ 'aniwe mat+a yaxeik+a xep+tehaka'erie	usted y su hijo se parecen mucho

wixárika	**español**
wani 'ukiratsi hepa+ p+tiyuxexeiya	juan parece viejo
wani ti'+kitame hepa+ p+ane	juan parece maestro
kepetitewa	¿comó te llamas?
pekuru nep+titewa	me llamo pedro
pux+ka	hace calor
puha+t+	hace frio
pu'eeka	hace viento
p+x+ka	hace sol
p+wiiye	está lluviendo
p+kawiiye	va a llover
titak+ paapa tiyutiwewiwa	¿con qué hacen las tortillas?
paapa xakwitsarik+ p+yutiwewiwa	las tortillas se hacen con nixtamal
titatsie reyani 'akie	¿comó se va a tu pueblo?
hakewa atsukari petuiya	¿dónde venden azucar?
mana wa+kawa p+tituiya	ahi venden muchas cosas
wani putah+awarie presidente municipalk+	nombraron a juan presidente municipal
wani kauka'iyari p+netiutah+awix+	juan me llamó tonto
ne p+kutsu	estoy durmiendo
pep+kutsu	estás durmiendo
p+kutsu	está durmiendo
pekutsu	está durmiendo
tep+kútsu	estamos durmiendo
xep+kútsu	ustedes están durmiendo
mep+kútsu	ellos están durmiendo
mepekútsu	ellos están durmiendo
p+ka kutsu	no esta durmiendo
wa+ka p+kakutsu	duerme poco
kwinie p+tikukutsu	duerme mucho
hek+ta pukukutsu	duerme de dia
hek+ta kwinie p+tikukutsu	duerme mucho de dia
y+wik+ta kwinie p+tikukutsu	duerme mucho de noche
y+wik+ta p+kakukutsu	no duerme de noche
heek+ta p+kakukutsu	no duerme de dia
nunutsi peuku	el niño se durmió
nunutsi peukuxime	el niño se está durmiendo
nunutsi peukuni	el niño se va a dormir
nunutsi hutarieka peuku	el niño se durmió de nuevo
nunutsi 'aix+ p+katiuku y+wik+ta	el niño no durmió bien anoche
nunutsi peuku t+ma kamiunitsie	el niño casi se durmió en el camión
kets+ m+kit+ hapa p+yehaune	el pez muerto está flotando en el agua
k+ye hapa p+yehaune	la madera flota en el agua
m+k+ yap+neti'uti wawiriwa paapa	ella siempre me pide tortillas
m+k+ paapa p+netsi'uta wawiri yumama hetsie mieme	ella me pidió tortillas para su madre
m+k+ paapa meti'uta wawiri	¿ella te pidió tortillas?
h+ m+k+ paapa p+netsi'uta wawiri	si,ella me pidió tortillas

wixárika	**español**
kem+'ane paapa metsi'uta wawiri	¿quién te pidió tortillas?
tita metiuta wawiri m+k+	¿qué te pidió ella?
m+k+ paapa p+kanetsi'uta wawiri	ella no me pidió tortillas
m+k+ heiwa paapa p+netsi'uti wawiriwa	ella a veces me pide tortillas
hawaik+ m+k+ paapa p+kanetsi hawawiriwani	ella nunca me pide tortillas
m+k+ waik+ paapa p+kanetsiuta wawiri	ella nunca me pidió tortillas
m+k+ 'ik+ kwikari p+netsi'uta +kit+a	ella me enseñó esta canción
m+k+ 'ik+ waikari p+netsi'uta '+kit+a	ella me explicó el juego
takai miratuiya nemetsihexei	te vi ayer en el mercado
takai nep+kamatsihexei m+ratuiya	ayer no te vi en el mercado
takai penerexei muratuiya	me vistes ayer en el mercado
kem+'ane pepexei takai m+ratuiya	¿a quién vistes ayer en el mercado?
tita perexei takai m+ratuiya	¿qué vistes ayer en el mercado?
hakewa kepauka pep+netsexei	¿dónde y cuándo me viste?
xupureru keneinanairieni p+nerah+awekai	me pidio que le comprara un sombrero para él
papaya xupureruya kenenanairieni p+nerah+awekai	me pidio que le comprara un sombrero para su padre
tateikie pep+mie m+k+ painekai	el ordenó que fueras a san andrés
m+k+ matiutan+'a tateikie	el te ordenó ir a san andres
m+k+ matsi'utaxanetax+ pem+reinawatsirik+	ella te acusó de haberle robado
tita petimate	¿qué sabes hacer?
tixa+ nep+katimate	yo no sé hacer nada
nep+karamate	yo no sé nada
perahauwe	¿sabes nadar?
teiwarik+ xeik+a wanieutaniuwe	juan soló sabe hablar español
teiwarik+ petiniuwe	¿sabes español?
teiwarik+ nep+niuwe	sé español
nep+kahahauwe	yo no sé nadar
kepauka pep+retima kename nenua	cuándo supistes que yo había llegado?
kepa+ pep+retima kename nenua	¿cómo supistes que yo había llegado?
pem+ramaikak+ nekametinah+aweni	te lo digo para que lo sepas
teiwari niukieya peretima	¿aprendistes español?
nekutsi kwikarik+ p+netiuti +kit+a	mi abuela me enseñó a cantar
nekutsi kwikarik+ p+kanetiuta'+kit+a	mi abuela no me enseñó a cantar
m+kati nawayanik+ neukiyari p+netiuti+kit+a	mi padre me enseñó a no robar
huye keneneuxeiyatsit+a	¡enséñame el camino!
amamatsie pereutu'u keneneuxeitsit+a	¡enséñame lo que tienes en la mano!
huyeta nep+tiwarexei meteyuwa+kawame	vi muchos animales en el camino
nepenierix+ kem+rey+ m+k+ kita	vi lo que pasó en la casa
axa+ta perekwakame nemetsihexei	vi que estabas comiendo solo
paapa nepeuyeh+wa	quiero tortillas
ximeri m+ratuiya nep+miekaku	quiero ir temprano al mercado
kepetiyurienikeyu	¿qué quieres hacer?
'ena nep+nehayewakeyu	quiero quedarme aquí
ya nep+tinaki'erie ena pem+kunuani	quiero que te quedes aquí
'ena axa+ta pemukunuani ya nep+katinaki'erie	no quiero que te quedes solo aquí

wixárika	**español**
wani kiena nemekunuani ya p+renaki'erie	juan queria que me quedara en su casa
waninemekunuani ya p+karenaki'eriekai	juan no quería que yo me quedara
manzana pep+netsiuta xat+a	me prometistes una manzana
manzana pep+netsiminikiekai pep+netiuxat+a	me prometiste que me ibas a dar una manzana
tateikie pep+netsi anuwit+nikekai pep+netiuxat+a	me prometistes llevarme a san andrés
yutsitsie petiutierie	¿crees en dios?
x+ari m+k+ wani mat+a pet+a	creo que él se fue con juan
wani pamie nep+karaeriwa	dudo de que venga juan
wani pamie nep+ra'eriwa	no dudo de que venga juan
yuri nep+katierie kem+tiuxa	dudo de su promesa
yuri nep+katierie kem+tiuxa	no creo en su promesa
yuri nep+ketierie m+ya m+nerah+awekai	no creo en lo que me dijo
kem+'ane mamie nep+imate	no sé quién viene
peramate kem+'ane m+kahamie	¿sabes quién no viene?
nep+karamate kem+'ane munua	no sé quién vino
nep+karamate kem+'ane mamie	no sé quién va a venir
nep+karamate hakewa meut+a	no sé a dónde va
nep+karamate titayari m+reukuyeix+a	no sé por qué vino
nep+karamate kem+rane	no sé de dónde vino
nep+karamate kepa+ itipari 'aix+ muyurieni	no sé cómo arreglar la puerta
nep+karamate tita wani m+titua	no sé qué vende juan
nep+karamaikai peamieme	yo no sabía que venías
nep+karetimai kem+'ane meukuyeix+a	no supe quién vino
nep+rat+mai nereunanikie	me olvidé de cerrar la puerta
nep+rat+mai 'ek+ pem+nuanikiekai	me olvidé de que llegabas hoy
kem+titewa nep+rat+mai	me olvidé su nombre
tita peret+mai	¿qué olvidaste?
m+k+ wikwinie p+rawiwe	ese pájaro puede volar mucho
m+k+ wiki p+kaawiwe	ese pajaro no puede volar
k+yetsie nep+kahanuti makiwe	no puedo subirme al árbol
kita nep+ka'utaha kitenie m+reunakaik+	no pude entrar en la casa porque la puerta estaba cerrada
ha tsimepenitsie pepanuyeiwe	vas a poder cruzar el río cuando tenga poca agua
nep+reuta ut+awexeik+a rapi nep+kahexeiya	sé escribir,pero no puedo porque no tengo lápiz
m+k+ k+yetsie nepeumakim+k+ xeik+a nep+kay+we.	quiero subirme a ese árbol pero no puedo
m+k+ k+yetsienepanutimakiwe xeik+a ye nep+kareunaki'	puedo subirme a ese árbol,pero no quiero
nep+tita 'uximayata	tengo que trabajar
takai tsinep+kareti 'uximayatax+	ayer tuve que trabajar mucho
patini 'uximayakam+k+	tienes que trabajar
xatsi p+taiyar+wa	hay que quemar la basura
xatsi kwanetataiyarieni	va a haber que quemar la basura
p+kanetsinake pemeyeikani	no me gusta que te vayas
kene 'ata'eriwat+ wa'at+ tep+teta 'uximayata	acuérdate de que tenemos que trabajar

wixárika	**español**
m+k+ te+teri mep+kanetsinake	no me gusta esa gente
m+k+ 'ikwai p+kanetsinakie	no me gusta esa comida
kene'a'eriwat+ kita pekaniuye 'itiem+k+	acuérdate de barrer la casa
p+netseta iwawiyax+ tita nem+reuye h+akai	me preguntó qué quería yo
net+riyama ya nep+tiwaku eriwa	siempre pienso en mi familia
ke'ane ts+ meuyeh+akaku	no sé cuál quiere
meri ya nep+rakewe	acostumbro levantarme temprano
'ikwai 'ena mieme nep+ka 'iyamate	no me acostumbro a la comida de aqui
meri m+kekanik+ pi'iyamatsit+a	lo acostumbró a levantarse temprano
hik+ nep+tayuani 'etsik+	hoy empiezo a sembrar
takai neputayua 'etsik+	ayer empecé a sembrar
'uxa'a 'etsik+ nep+tayuani	mañana voy a empezar a sembrar
hik+ 'ix+arari p+tayuani kiekaritsie	hoy empieza la fiesta en el pueblo
hik+ nep+tita heyewa 'etsik+	hoy acabo de sembrar
takai nep+tiuta hayewax+ 'estik+	ayer acabé de sembrar
'uxa'a nep+tita hayewa 'etsik+	mañana voy a acabar de sembrar
hik+ nepiyeweiyani 'etsik+	hoy voy a seguir sembrando
takai nepiyeweiyakai 'etsik+	ayer segui sembrando
huye keneuweiyani	¡siga el camino!
kiena nepeiku ix+ari	fui a verlo a su casa
heiwa tateikie pekaranuyeiwe	¿has ido alguna vez a san andrés?
titayari xeme xekatehanuk+	¿por qué no fueron ustedes?
ya p+tiuta 'axe kawayutsixi wawauriyarik+	viene todos los dias a buscar los caballos
kenanutimaki yeuka kenaka'in+i	súbete a bajar ese aguacate
niwa kenenaparewimie	ven a ayudarme
niwa kenenaparewimie puritu 'ikat+arikak+	ven a ayudarme a cargar el burro
neniwema ki h+iyarik+ mepukunuax+a	mis hijos se quedaron a cuidar la casa
kenanukayaka ketinatipini	¡baja a recogerlo!
kita 'ayeneka p+netsihetah+awix+	salió de la casa a saludarme
m+ratuiya nep+yemiexime	estoy por ir al mercado
m+ratuiya nep+yemieximekai m+ netiukunua	estaba por ir al mercado,pero me quedé en casa
hik+ nep+titi 'uximayatat+yani	hoy me pongo a trabajar
kiriwa 'ena p+h+a	dejó aquí la canasta
kitenie reutenime puhayewax+	dejó abierta la puerta
keneupit+a panutayani	déjalo entrar
tikuyet+t+ 'uximayatsikak+ p+katiuhayewax+	no déjo de trabajar,aunque estaba enfermo
ya p+tiuti 'uximayatax+	trabajó todos los días
xei witari panuyemie nekati 'uximayat+	hace un año que no trabajo
takai pai nep+kati 'uximaya	no trabajo desde ayer
hik+ waitari wa+kawa nep+tiuti 'uximayatax+	este año trabajé mucho
takai pai nep+ti 'uximaya	estoy trabajando desde ayer
ximeri wa+kawa nep+tiuti 'uximayatax+	esta mañana trabajé mucho
taikai ya nep+tiuti 'uximayata	siempre trabajo de trade
'uxa'a nep+tita 'uximayata	mañana trabajo
petita 'uximayata	¿trabajarás?
patinita 'uximayatam+k+	¿trabajarás mañana?

wixárika	**español**
tawari hik+ nep+tita uximayata	hoy voy a trabajar otra vez
wani p+ti'uximatari	juan ya está trabajando
ketine 'uximayakari	¡empieza a trabajar!
m+k+ k+ye kwit+ patiweni	ese arbol se va a caer pronto
nepenekatewiyat+ya nem+ka'i parewik+	me dio pena no poder ayudarlo
m+k+ k+ye kwit+ patiweni	ese árbol se va a caer pronto
k+ye patiwe	el árbol se cayó
ne'+ka putahai	se me hinchó el pie
ne'+ka p+ha	tengo el pie hinchado
nekamixa neputahaxuma	me ensucié la camisa con lodo
nekamixa putahaxumarix+	se me ensució la camisa con lodo
tai 'aix+ p+tiutatawe	el fuego arde bien
'aki p+ta'a	tu casa está ardiendo
paapa neputix+tsit+a	quemé las tortillas
paapa putix+	las tortillas se quemaron
paapa p+tata'i	las tortillas están quemadas
nemu'u p+netsi'u kukwine	me duele la cabeza
neniwe p+tiuta kwinix+	mi hijo se enfermó
neniwe p+tikuye	mi hijo está enfermo
nep+tikuye	estoy enfermo
nep+tikuyekai	estuve enfermo
netei pum+	mi tía se murió
netei pem+	mi tía se murió
m+k+ 'uki p+m+ki	ese hombre está muerto
wani kwinie p+rayunaanaiwie	juan se ríe mucho
wani matsi 'atse	juan se ríe de ti
m+k+ 'uki takai pemierie	a ese hombre lo mataron ayer
teik+mana nepe'uxix+	me cansé en la subida
hik+ nepeu'uxe	ahora estoy cansado
kawayu pe'uxix+	el caballo se cansó
kawayu peu'uxe	el caballo está cansado
kawayu pe'uxit+a	cansó el caballo
mariya nunutsi pukut+a	maria está durmiendo al nene
mariya peuku	maria se durmió
mariya peukut+a nunutsi	maria hizo dormir al nene
wani ki 'ap+tapariya	juan va a agrandar la casa
wani ki tsip+taperiya	juan va a achicar la casa
wani ki 'ap+tapariya xeiwiyari hanuyeyeikakaku	juan agranda la casa todos los años
kukuri putaxetare x+ka 'utawani	el chile se pone rojo cuando se madura
wani ki putaturiyax+	juan blanqueó la casa
xari p+y+xa+ye	la olla es negra
k+tsi xari putay+xariyax+	el humo ennegreció la olla
'ik+ nunutsi 'axa p+tiuka 'iyari peru 'aix+ tiuka 'i	este niño ahora es malo pero se va a hacer bueno
wani p+netsi'uta nanait+a	juan me hizo reír
wani p+netsi'uti nanait+a	juan me hizo reír
mexa nepeuxawariyax+	agujereé la tabla

wixárika	**español**
mexa peuxawa	la tabla está agujereda
tsik+iwiti neputahaxuma	enlodé la canasta
tsik+iwiti p+haxuma	la canasta está enlodado
tsik+iwiti putahaxumarix+	la canasta se enlodó
tsik+iwiti wiwierieya p+haxuma	la canasta tiene lodo en el asa
tsik+iwiti wiwieri p+haxuma	el asa de la canasta está enlodado
nets+k+ pem+	mi perro se murió
nets+k+ pum+	mi perro se murió
wani nets+k+ pemi	juan mató a mi perro
wani nets+k+ pumi	juan mató a mi perro
wani 'utay+k+ nets+k+ pumierie	juan hizo matar a mi perro
m+k+ 'utay+ku nets+k+ nepumi	el me hizo matar a mi perro
wa+riyarika m+k+ netsi 'ait+akaku ts+k+ nepumi	el me obligó matar a mi perro
ne nepekuk+'ai	yo traje la leña
p+netsi k+'aitsit+a	me hizo traer la leña
mariya ha p+x+riyax+	maria calentó el agua
ha p+x+ka	el agua está caliente
m+k+ 'ixuriki pep+ka'itsiturariyani	¡no arruges ese trapo!
ixuriki p+itsiturie	el trapo está arrugado
mariya xak+r+te putihauxi	maria lavó los platos
xak+r+te p+hauxiniet+ka	los platos están lavados
wani kitenie p+reuna	juan cerró la puerta
wani kitenie p+reuyepi	juan abrió la puerta
kitenie p+reunarix+	la puerta se cerró
kitenie p+reuyepierix+	la puerta se abrió
kitenie p+reuname	la puerta está cerrada
kitenie p+reuyepie	la puerta está abierta
kawayu p+nautsane	el caballo está corriendo
wani kawayu p+nautsit+ane	juan está haciendo correr al caballo
wani putaya	juan se sentó
wani nunutsi mexatsie paye	juan sentó al niño en la mesa
'ik+ waxa neniwema mepika'e	esta milpa fue sembrada por mis hijos
'uta kwit+ paye'a	llegó cantando
'ukwikar+met+ punua	llegó cantando
'uta wik+ayat+ panuyet+a	pasó silbando
kanetsiha xexeiyat+ panuyet+a	pasó sin verme
'uxet+ p+nua	llegó cansado
nanaik+ patiwe	se río hasta caerse
ya/kareu naki'eriet+ p+ti'uximaya	trabaja sin ganas
t+r+ka+yemek+ p+ti'uximaya	trabaja con ganas
yu'+kama pu+nua	vino a pie
kawayutsie hakait+ punua	vino a caballo
xupureru hanaket+ pet+a	salió con sombrero
xupureru kahanaket+ pet+a	salió sin sombrero
mex+iwa p+ti'uximaya	trabaja apurado

wixárika	**español**
penakuh+ake	lo amarró apenas
wa+riyarika penakuh+a	lo amarró con dificultad
p+netsiutaiwi	me llamó a gritos
kwi keneutaniu	¡habla en voz alta!
an+ari keneutaniu	¡habla en voz baja!
kaunari tiraunime kenawieka	¡mantén el mecate tirante!
kaiwat+ kaunari kenawieka	mantén el mecate firme
amama kenexutseriyani	¡manten el brazo rigido!
'amama keneweraniyani	¡mantén el brazo flojo!
heitserie keneyeye'a	¡camina en linea recta!
heitserie keneutayexi	¡siéntate derecho!
tete peh+a kayut+r+karima	arrojó la piedra sin fuerza
tete peh+a karima	arrojó la piedra con fuerza
xuya tet+atapai peuke	la espina se me clavó profundamente
xuya herie peuke	la espina se me clavó superficialmente
kenat+mina 'an+ari maxitekietsie	¡frota la herida con suavidad!
kanat+mina t+r+karima maxitekietsie	¡frota la energia con energia!
kuxitari kenanukuhani'an+ari	¡levanta el costal poco a poco!
kuxitari kenanukuhani xeimieme	¡levanta el costal de una vez!
karima pep+kareunar+mani	no cierre la puerta de golpe
hek+ta xeiya nep+yeiwe	de dia puedo caminar rápido
y+wik+ta 'an+ari nekaniumiem+k+	de noche tengo que caminar despacio
wani kaunarik+ pay+h+aya	juan usa un mecate como cinturón
titak+ tium+ atatatsi	¿de qué murió tu tió?
tsipurikiyak+ pum+	murió de viruela
'ik+ tsik+iwiti kerayeaxe	¿cuánto cuesta esta canasta?
xei'in+ariyari payeaxe	cuesta un peso
kerahete 'ik+ kuxitari	¿cuánto pesa este costal?
tamamate kiruyari p+rahete	pesa diez kilos
kaunari nawaxak+ kenanuxiteki	¡corta el mecate con el cuchillo!
kenanutsanaamamak+	¡rómpelo con las manos!
keneutah+a 'ik+ kaunarik+	¡atalo con este mecate!
kepeti+ni k+yexi	¿comó vas a llevar la leña?
puxutsie nepitituani	la voy a llevar con el burro
titatsie pereyet+a kiekaritsie	¿en qué viniste del pueblo?
kamiunitsie nepeyet+a (kiekaritsie)	vine en camión (del pueblo)
muratsie nepeyet+a	vine en mula
tsik+iwitite tak+k+ p+yuwewiwa	las canastas se hacen de palma
xarite haxuk+ p+yuwewiwa	las ollas se hacen de barro
xama hukaiwa keneutaketsina	¡aparta la rama con el pie!
xari k+yek+ pep+kakuwayani	¡no golpees la olla con el palo!
x+ka x+nariya wewim+k+ni haxu'+xa kanin+it+ariwani	para hecer adobe se mezcla lodo con paja
wani mat+a p+nua	vino con juan
nehamiku mat+a punua	vino con mi amigo
nehamat+a p+nua	vino conmigo

wixárika	**español**
wani mat+a pekuru uxeik+a mep+teutiuximayata	juan y pedro trabajan juntos
kepa+mex+a pepekuyeix+a	¡cuántas veces vinistes?
hekewa paka nunutsi	¿dónde está sentado el niño?
titatsie raka nunutsi	¿dónde está sentado el niño?
neh+xie paka	está sentado frente de mi
newarita paka	está sentado detrás de mi
newarita paka	está sentado adelante de mi
ne'aurie paka	está sentado junto a mi
nepapa kita peutaha	mi papá entró en la casa
hekewa peuha apapa	¿a dónde entró tu papá?
nepapa kita p+wayet+a	mi papá salió de la casa
hakewa apapa pewayet+a	¿de dónde salió tu papá?
mitsu k+yetsie panutimakix+	el gato se subió al árbol
mitsu k+yetsie pakamakix+	el gato bajó del árbol
yupurutsata tumini peukamanax+	puso el dinero en el bolsillo
yupurutsata tumini pati'+i	sacó el dinero del bolsillo
wani kitenie peuyemie	juan salió por la puerta
wani xawata peukawe	juan se cayó en el pozo
wani +paritsie paya	juan se sentó en la silla
wani kwiepa putaya	juan se sentó en el suelo
wani kitsie pakawe	juan se cayó del techo
wani kitsie pukawe	juan se cayó sobre el techo
wani oaxaca peyet+a	juan viene de oaxaca
wani kiekaritsie peyet+a	juan viene del pueblo
wani wxata peyet+a	juan viene de la milpa
wani uyeta p+kaheyet+a waxata p+ta	juan no vino por el camino si no a través de la milpa
wani hat+a peukawe	juan se cayó en el río
wani hat+a peukatsunax+	juan se tiró al río
wani hat+a p++wane	juan se está bañando en el río
hat+a pai tekanihuni	vayamos hasta el río
wani nekie pait+ ye'aka p+ka'utay+ne x+ari m+ye yaan	juan llegó hasta mi casa y no quiso seguir caminando
wani nekie peyeikakai	juan anduvo por mi casa
wani hat+a hepa+tsie peyet+a	juan vino por el lado del rio
huye hix+apa tete kwinie net+ puka	en medio del camino hay una piedra grande
huye tetsita k+yexi peut+ka	al lado del camino hay árboles
xat+ hix+apa patari	el centro del comal está quebrado
xat+ hix+apa xaip+ paka	hay una mosca en el centro del comal
xat+ hix+apa putarix+	el comal se partió por enmedio
wani hix+apa mayewe p+h+k+	juan es el de enmedio
'uwa p+ta tep+kaeni	vamos a sembrar de este lado de la casa
xeme xete'ukaetsa huye tetsita	¿siembran ustedes en la orilla del camino?
'uki panuyet+a	el hombre salió
'uki p+tawe	el hombre está borracho
uki m+wayet+a p+tawekai	el hombre que salió estaba borracho
iki tawet+ p+wayet+a	el hombre salió borracho

wixárika	**español**
uki m+tawekai nepexei	vi al hombre que estaba borracho
uki tawekame nepetaxei	encontré borracho al hombre
uki m+tawekai mat+a nepunua	vine con el hombre que estaba borracho
kem+'ane 'uki m+tawekai p+wayet+a	¿cuál de los hombres que estaban borrachos salió
uki tsik+ pumi	el hombre mató al perro
ts+k+ uki p+k+kewekai	el perro estaba mordiendo al hombre
ts+k+ p+netsikukewekai	el perro me estaba mordiendo
ts+k+ putikuyekai	el perro estaba enfermo
tsik+ m+k+'uki mumi p+tikuyekai	el perro que mató ese hombre estaba enfermo
tsik+ m+k+'uki p+netsikukewekai	el perro que mató ese hombre me estaba mordiendo
kem+'ane panuyet+a ts+k+ memumi	¿cuál de los hombres que mataron al perro salió?
'uki panuyet+a ts+k+ nemimi'iri	salió el hombre al que le maté el perro
'uki p+wayet+a kita ts+k+ m+wami	el hombre salió de la casa en la que mató al perro
kutsira nepekaxei pemenuh+akaitsie	encontré el machete donde lo dejastes
turukitsie nepeyet+a temetayeix+atsiepai	vine en camión desde donde nos separamos
puteyu kenenahanit+a ha mayema	dame la botella que tiene agua
puteyu keneneuhanit+a ha hayemakame	dame una botella que tenga agua
caja keneneu+it+a cerillo pematiuta	dame la caja donde guardas los cerillos
nawaxa keneneukweit+a k+manak+ wai pemutixi-tekie	dame el cuchillo con que cortas la carne
tewi nep+kaimate hamat+ana pemunua	no conozco al señor con el que viniste
tewi nepexei wakanaripemeituiri	vi al señor al que le vendistes las gallinas
'ik+ 'ukiratsi temexei he'eneme p+h+k+	este es el señor que vimos sembrando
m+k+ p+'ukiratsi	el es viejo
m+k+ wa+ka matsi p+'ukiratsi	el es más viejo
wa+ka m+'ukiratsi pum+	el más viejo se murió
xat+ panakatari	el borde del comal está quebrado
wani kawayaya pum+	el caballo de juan se murió
ahamiku kawayaya pum+	el caballo de tu amigo se murió
ki yuta+ta peuka 'unix+	el techo de la casa se cayó
nepapa kiya yuta+ta peuka'unix+	el techo de la casa de mi padre se cayó
kutsira kwaxieya panamuri	el mango del machete está quebrada
ha 'uwa mieme 'axa pa'ane	el agua de este pozo es mala
ha hat+a mieme 'axa p+'ane	el agua del rio está sucia
nawi h+iyame nepexeiya	tengo un cinturón de piel
xarita ha p+kwana	el agua de la olla está hirviendo
'eeka hukaiwa mieme tsip+katiha+t+	el viento del norte es frío
kape kuxitariyari paxawa	el costal para el café esta agujereado
m+k+ kape kuxitariyari kwinie p+rahete	ese costal de café pesa mucho
tsik+ meuy+xa+ye p+netewa	el perro negro es mío
ts+k+ tsim+pe meuy+xa+ye p+netewa	el perro chico negro es mío
x+nariya m+kwakwaxi tekiyari kwinie p+titse'i m+yuyu	la pared de ladrillos es más fuerte que la de adobe
m+k+ nereuyeta p+teewi	él es más alto que yo
m+k+ aniwe reuyeta p+teewi	él es más alto que tu hijo
m+k+ 'aputeewi nehepa+	él es tan alto como yo

wixárika	**español**
m+k+ 'aputeewi aniwehepa+	él es tan alto como tu hijo
m+k+ meri punua ne'arikeke	él vino antes que yo
m+k+ meri punua 'aniwe arikeke	él vino antes que tu hijo
ne meri nepunua m+k+ arike	él vino mas tarde que yo
'aniwe meri punua m+k+'arikeke	él vino mas tarde que tu hijo
m+k+ yaak+ p+rakumex+a ne kwinie	él es menos rápido que yo
m+k+ yaak+ p+rakumex+a aniwe kwinie	él es menos rápido que tu hijo
m+k+ 'aix+ p+tiuuximayata ne hepa+	el trabaja tan bien como yo
wakanari memeutuxa yunaime nep+watinaneni t+ma	voy a comprar casi todas las gallinas blancas
hipat+ m+k+ t+ri mep+tekukuye	algunos de esos niños están enfermo
yunait+ etsiwat+kaku mep+tekukuye m+k+ t+ri	casi todos esos noños están enfermos
xewit+t+ma nunutsi p+katikuye	ninguno de los niños está enfermo
'ik+ t++ri ruritse yuxexuit+ xexuime meputikwai	cada uno de estos niños comió un dulce
'ik+ t++ri xewit+ matsitah+awe	cualquiera de estos niños te llamará
m+k+ t++ri yuwa+kawa mep+tekukuye	muchos de estos niños están enfermos
m+k+ t++ri yameyupa+met+ mep+tekukuye	pocos de estos niños están enfermos
kwiniw mey+pa+met+ t+ri mep+tekukuye	bastantes niños estan enfermos
wa+kawa tsarape nepetua	vendí muchos sarapes
yapa+ nepetua tsarape	vendí algunos sarapes
tsarape tsipa+meme nepetua	vendí pocos sarapes
wa+kawa tsarape nepetua	vendí vastante sarapes
'ik+ ikwai wa+kawa pukwai	comió demasiado de esta comida
kukuri wa+kawa putikwai	comió demasiados chiles
xewi	uno
huta	dos
haika	tres
nauka	cuatro
'auxuwi	cinco
'ataxewi	seis
'atahuta	siete
'atahaika	ocho
'atanauka	nueve
tamamata	diez
xeitewiyari	veinte
xeitsientuyari	cien
xeime kawayu nep+tewa	tengo un caballo
turutsixi meyuhutame meheuy+y+wimenep+watewa	tengo dos bueyes negros
kiena xeimieme nepanut+a	fui a su casa una vez
kiena hutak+a nepanut+a	fui a su casa dos veces
kiena haikak+a nepanut+a	fui a su casa tres veces
ukitsi yuxexuit+ kita mepayeneikakai	los hombres salieron de la casa uno por uno
'ukitsi meyuhutatat+ kita mepayeneikakai	los hombres salieron de la casa de dos en dos
'ukitsi meyuhaikakat+ kita mepayeneikakai	los hombres salieron de la casa de tres en tres
tetexi xexuime xekeneutit+kix+a	¡agarren una piedra cada uno!

wixárika	**español**
k+yexitsie xeketena 'ut+ax+a huhutakame xehanayehaye	hagan una señal cada dos árboles
wani pet+a pe'ixeiyak+	juan se fue para que nolo vieras
wani punua pemixeiyak+	juan vino para que lo vieras
kaxeta nepenanai k+yexi neikatam+t+	compré la carreta para llevar la leña
kaxeta nepunanai k+yexi pemanuikatak+	compré la carreta para que lleves la leña
tsepa m+wiye nep+yemie	voy a ir aunque esté lluviendo
tsepa m+tiwiye nep+ye.mie	voy a ir aunque llueva
pex+ka yemieni keneaku mex+it+wani	si vas a ir apúrate
nex+kaixeiya kutsira nepitawawirieni	si lo veo le pido el machete
nex+kaixeiyake kutsira nepitawawirienike	si lo viera le pidiría el machete
nex+kaixeiya ke kutsira nenitawawirienike	si lo hubiera visto le habría pedido el machete
nepet+a nem+kati'uximayak+	me voy porque no estoy trabajando
y+xa+ta p+kaumie m+mak+	no va solo por que tiene miedo
wani pukutsukai pem+nuatsie	juan estaba durmiendo cuando llegaste
wani pukutsukai pemunuatsieke panutanierix+	juan estuvo durmiendo hasta cuando llegaste
wani pukutsukai pekanuawekaku	juan estuvo durmido antes de que llegaras
wani peukuni penuayu	juan va a dormir cuando llegues
wani peukuni pem+nuanitsieke	juan va a dormir hasta cuando llegues
wani peukunike kepauka pem+nuani	juan va a dormir apenas llegues
'axa pep+kanetsi mat+aka nem+tikwakatsie	no me molestes cuando estoy comiendo
'axa pep+kanetsiutimat+wani nem+tiutikwa'akatsie	no me molestes cuando como
x+ka meheuk+nikuni kemehek+ne	si quieren irse que se vayan
wani kukuri p+kaiwieni neta hepa+natsiere	juan va a sembrar chile y yo también
wani kukuri p+kaiwieni ne waik+	juan va a sembrar chile pero yo no
wani p+kakaeni ne h+rix+a	juan no va a sembrar, pero yo si
m+k+ nep+kareuyeh+wa	no quiero nada de eso
tixa+t+ 'ena pereuyeh+wa	¿quiero algo de esto?
nexa+ta nep+kauyeiwe	no puedo caminar solo
ximeri xeik+a nem+tita uximayata nep+y+we	ya sólo puedo trabajar en la mañana
manari nep+mie	ya mero voy
tem+te uximayatayukai hik+ri tsip+katiutaikai	ahora es tarde para ir a trabajar
hik+ 'akuxi tsip+ka'upaukwa tem+te'uximayatayukai	ahora es temprano para ir a trabajar
'ik+ ke'uxa p+kwaiwa	esta planta es buena para comer
'ik+ tupiriya 'aix+ p+tiuaye m+k+ kwiniyarik+	esta planta es buena para esa enfermedad
'ik+ tupiriya p+ka'uaye m+k+ kwiniyarik+	esta planta es mala para esa enfermedad
wani tsip+katiu yeiwe	juan es rápido para caminar
'ik+ kwie tsip+karayetse'i p+kayu 'etsin+a	está tierra es demasiado dura para sembrar
hik+ri maxatsi 'axa tep+tewaxeiya h+ritsie	ahora es raro ver venados en el monte
m+k+ kawayu p+kayu maxiutsin+a	ese caballo es difícil de amansar
kwit+ yeiya m+t+namienike	ojalá venga pronto
x+arits+ kwit+ kanamieni	puede que venga pronto
tixa+ yapa+ xeik+a	no,deme menos
waika pereuyeh+wa	¿quiero mucho?

wixárika	**español**
tixa+ yak+ pa+ xeik+a	no,dene poco
kena yak+ pa+ xeik+a	no,dene poco