Universidad Nacional Autónoma de México

Posgrado en Ciencia e Ingeniería de la Computación
Lingüística computacional

EXTRACCIÓN LÉXICA BILINGÜE AUTOMÁTICA PARA LENGUAS DE BAJOS RECURSOS DIGITALES

T E S I S

QUE PARA OPTAR POR EL GRADO DE
DOCTORA EN CIENCIAS (COMPUTACIÓN)

P R E S E N T A

MARÍA XIMENA GUTIÉRREZ VASQUES

TUTOR: DR. GERARDO SIERRA MARTÍNEZ

CO-TUTOR: DR. ALFONSO MEDINA URREA
Posgrado en Ciencia e Ingeniería de la Computación

CIUDAD UNIVERSITARIA, CDMX, AGOSTO 2018

Paw dice que por la rudeza de la lengua mexicana no ha habido hasta ahora un español que pueda pronunciarla y que por la incapacidad de los mexicanos ninguno de ellos ha aprendido hasta ahora la lengua española; pero lo uno y lo otro distan mucho de la verdad.

Francisco Xavier Clavixero, *Historia antigua de México*

La expresión caos determinista es, en cierta forma, una contradicción en si misma, pero ilustra un tipo de comportamiento muy atractivo, equidistante del caos absoluto y del determinismo absoluto. Justo donde suceden las cosas interesantes. Es un comportamiento irregular pero sigue reglas definidas. Parcialmente predecible, parcialmente impredecible, pero con normas y estructuras subyacentes. Como la vida misma.

Alberto Pérez Izquierdo, *La teoría del caos: las leyes de lo impredecible*

A mis abuelos de la Ciudad de México, Elvia y Ricardo.

...y a mis abuelos del campo, Dominga y Don Beto, de los cuales sigo aprendiendo palabras en lengua Pame, esas que recuerdan de su juventud en San Luis Potosí.

Resumen

Esta tesis tiene como objetivo realizar extracción léxica bilingüe automática para lenguas que enfrentan una escasez de corpus paralelos y de recursos digitales. Nos enfocamos en el español-náhuatl, un caso de estudio con diversas particularidades: las lenguas son tipológicamente distantes, poseen gran riqueza morfológica, existe escasez de recursos digitales y tecnologías para el náhuatl así como una gran variación dialectal y ortográfica. Gran parte de los métodos tradicionales en PLN no trabajan bien bajo estas condiciones experimentales. En primer lugar, proponemos la creación de un corpus paralelo español-náhuatl digital a partir de fuentes físicas (libros). Posteriormente, proponemos utilizar representaciones que tomen en cuenta la morfología de las lenguas para facilitar la tarea de extracción léxica bilingüe entre una lengua altamente aglutinante y una lengua fusional. Proponemos también la combinación de diferentes aproximaciones estadísticas para estimar las correspondencias entre las dos lenguas a partir de un corpus paralelo pequeño.

En particular, combinamos modelos estadísticos de tipo estimativo y asociativo para extraer pares de traducción y generar así un lexicón semilla. Asimismo, a través de un grafo que modela las relaciones bilingües entre palabras, generamos representaciones vectoriales multilingües para cada palabra del español y del náhuatl. Finalmente, utilizando el lexicón semilla y las representaciones vectoriales obtenidas, aprendemos una transformación o mapeo lineal entre los vectores de las dos lenguas, de tal manera que se pudieran traducir palabras del español a náhuatl usando esta transformación.

Nuestra metodología muestra una mejora significativa en la extracción automática de pares de traducción, tomando como referencia el desempeño de los métodos y representaciones vectoriales populares en PLN cuya calidad es altamente dependiente de grandes cantidades de texto.

Abstract

The aim of this thesis is to perform bilingual lexicon extraction for languages that face parallel corpora and digital resources scarcity. We focus on Spanish-Nahuatl, this is a case of study with several particularities: the languages are typologically distant, they are morphologically rich, Nahuatl do not have many digital resources, moreover, this language has an important dialectal and orthographic variation. Under these conditions, traditional NLP methods do not have the best performance. We propose the creation of a digital parallel corpus based on physical sources (books). We use morphological based representations in order to facilitate the task of finding lexical correspondences between a highly agglutinative language and a fusional one. We propose the combination of different statistical approaches in order to estimate de bilingual correspondences, based on a small parallel corpus.

In particular, we combined estimation and association statistical models for extracting translations pairs and inducing a seed lexicon. We also generated multilingual word vector representations, for Spanish and Nahuatl, using a graph structure that model the bilingual relationships among words. Finally, we used the seed lexicon and the word vector representations for learning a linear map between languages. In this way we are able to translate Spanish words into Nahuatl.

Our methodology outperforms popular NLP methods for bilingual lexicon extraction and word vector representations. The quality of these approaches usually relies on the availability of big amounts of text.

Agradecimientos

A todos aquellos que hacen posible que la Universidad Nacional Autónoma de México (UNAM) y el CONACYT nos brinden la oportunidad de realizar un posgrado. Los que recibimos este privilegio, tenemos la responsabilidad de fomentar las condiciones para que en el futuro muchos más jóvenes tengan la misma oportunidad que nos fue concedida.

En particular agradezco a la beca Conacyt 370713 y a los proyectos CB-2012/178248 y 2016- 01-2225. Así como al programa de apoyos PAEP, brindado por la UNAM, que me permitió presentar mi trabajo en diversos foros nacionales e internacionales.

Agradezco a mis tutores, Alfonso Medina y Gerardo Sierra, por apoyarme y guiarme en mi inquietud de llevar a cabo este tema de investigación. Así como a los investigadores, miembros de mi comité tutoral y jurado, que dedicaron su tiempo en auxiliarme a mejorar mi tesis en diversas etapas de su desarrollo: Ivan Meza, Sofía Galicia y Grigori Sidorov.

Al investigador Polo Valiñas por permitirme cursar sus, invaluables, clases de náhuatl clásico y morfofonología del español. Al investigador Marc Thouvenot por compartirme su experiencia en el procesamiento computacional del náhuatl y por recibirme en su casa en Francia. También agradezco a mis profesores del CELE (ahora ENALLT) por ser mi primer puente hacia la lengua-cultura náhuatl.

Agradezco también a los diferentes estudiantes que mostraron interés en el proyecto *Axolotl* y que, durante su servicio social, ayudaron a enriquecerlo. Gracias al equipo administrativo del Posgrado en el IIMAS, Lulú, Amalia y Ceci, por ayudarme estos años a realizar los diversos trámites que me han llevado a culminar el doctorado.

Durante mis años de doctorado, tuve la gran fortuna de coincidir en el Grupo de Ingeniería Lingüística con excepcionales colegas de diferentes áreas del conocimiento. Esta rica diversidad de perfiles generó interminables discusiones en, lo que bautizamos como, *el café académico*. Les expreso mi gratitud pues de alguna forma mi trabajo está in-

fluenciado por ustedes: Ignacio Arroyo, Carlos Morales, Carlos Méndez, Rocio Cerbón, Julián Solorzano, Octavio Sanchez. De manera muy especial agradezco a Víctor Mijangos, lingüista matemático, y entrañable amigo, que a través de innumerables pláticas y colaboraciones sembró en mi una genuina admiración por la labor lingüística.

Desde luego agradezco mucho a mi familia y a mis amigos. A Emilio, María, Emilio Ricardo y Rodrigo, por todo el apoyo y por ser los mejores compañeros de viaje que la vida pudo darme.

A Javier Santillán por su siempre breve pero importante paso por mi vida. Gracias por invitarme a ser parte de *Elotl*.

Finalmente, a todos aquellos que en algún punto de este camino nos hemos cruzado y me han dado apoyo o inspiración para mi labor, que me han alentado a creer que se puede, y que tiene sentido, crear tecnología para las lenguas habladas en México

Namechtlasohkamachililia
Se los agradezco mucho

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Este primer capítulo contiene una breve introducción al área de estudio en la que se enmarca esta tesis. Se describe de manera general la problemática abordada, así como los objetivos, la hipótesis y preguntas de investigación que busca resolver el presente trabajo.

1.1. Planteamiento del problema

El procesamiento del lenguaje natural (PLN), también conocido como lingüística computacional, es una área multidisciplinaria que se encarga del estudio científico del lenguaje humano bajo una perspectiva computacional, así como del desarrollo de tecnologías del lenguaje. Entre los muchos temas de interés para esta área se encuentran aquellos relacionados con multilingüismo, es decir, poder tratar mediante métodos computacionales más de un lengua.

Un recurso frecuentemente explotado en PLN son los corpus textuales, colecciones de textos representativas del uso de la lengua que permiten estudiar y modelar diversos fenómenos del lenguaje humano. En particular, para cuestiones multilingües, se utiliza el corpus paralelo, que es un tipo de corpus textual formado por documentos en una lengua fuente y su respectiva traducción en una o más lenguas destino.

Los corpus paralelos constituyen un recurso muy valioso debido a la información léxica bilingüe que contienen, permiten el desarrollo de diversas tecnologías del lenguaje, por ejemplo, la construcción automática de lexicones bilingües y los sistemas estadísticos de traducción automática.

Nos enfocaremos en la tarea de extracción léxica bilingüe, que consiste en obtener automáticamente pares de traducción a nivel palabra a partir de un corpus. Esta tarea permite la construcción de lexicones bilingües de manera automática, lo cual resulta útil para diversos fines. Por un lado, los diccionarios bilingües son recursos costosos que necesitan la intervención manual de especialistas; no siempre son recursos disponibles para todas las lenguas, especialmente cuando una o más de las lenguas tratadas son de bajos recursos digitales. Por otro lado, encontrar correspondencias a nivel palabra puede ser útil para algunos tipos de sistemas estadísticos de traducción automática que requieren de tablas de traducción a nivel palabras o frases, para estimar las traducciones de unidades más grandes.

La tarea de obtener pares de traducción a partir de un corpus ha sido un área activa de investigación desde hace varios años, especialmente a partir de la existencia de grandes cantidades de corpus paralelos digitales que permiten modelar las relaciones entre las unidades léxicas de los documentos paralelos.

Las primeras aproximaciones para realizar extracción léxica bilingüe están estrechamente relacionadas con la tarea de traducción automática. En la década de los 90s, el área de traducción automática gozó de un interés renovado; se establecieron los modelos de traducción estadística que están basados en nociones de teoría de la información y que permiten estimar la probabilidad de que una oración sea traducción de otra (basados en la existencia de grandes corpus paralelos) (Brown et al., 1993). Estos modelos de traducción estiman primero la alineación entre las palabras (o pequeñas frases) de las dos lenguas. Estas distribuciones de probabilidad de traducción a nivel palabra, también llamadas tablas de traducción léxica, permiten extraer léxico bilingüe.

Sin embargo, la calidad de esta y otras aproximaciones populares es altamente dependiente de la cantidad de datos. Se requieren grandes cantidades de corpus paralelos para que un sistema tradicional de traducción automática o de extracción léxica bilingüe, funcione adecuadamente.

Como no todas las lenguas poseen grandes cantidades de texto paralelo digital, procesado,

y listo para usarse, es necesario proponer métodos adicionales o alternativos para encontrar las correspondencias léxicas bilingües. Esta situación de escasez se acentúa aún más cuando tratamos con lenguas de bajos recursos digitales. Esto es, lenguas que no poseen, por diversos motivos, una abundante producción de textos digitales así como herramientas computacionales que ayuden a su procesamiento. Esta escasez de datos convierte a este tipo de lenguas en un reto de investigación para diversas disciplinas, incluido el PLN.

Gran parte de los estudios y desarrollo de tecnologías del lenguaje, se concentra en un subconjunto reducido de lenguas, por ejemplo el inglés, para el cual existe una vasta cantidad de recursos disponibles.

Además de la escasez de recursos, otro factor que dificulta la tarea de extracción léxica bilingüe es tratar pares de lenguas que son distantes. Esto es, lenguas que pertenecen a diferentes familias lingüísticas y que no comparten similitudes a nivel ortográfico, morfológico, sintáctico, etc.

Nuestro caso de estudio se enfoca en el par de lenguas español-náhuatl, dos lenguas habladas en el mismo país, México, pero pertenecientes a diferentes familias lingüísticas (indoeuropea y yuto-nahua). Estas lenguas no tienen una producción equiparable de textos, en el caso del náhuatl hay escasez tanto de corpus digitales monolingües como de paralelos. En contraste, el español es una de las lenguas con más hablantes en el mundo, posee una amplia producción de documentos digitales y tiene disponibles diferentes tecnologías del lenguaje.

Al ser de familias lingüísticas diferentes, estas lenguas exhiben fenómenos muy distintos, particularmente en la morfología (estructura interna de las palabras). El náhuatl es una lengua de morfología aglutinante con tendencia polisintética, esto significa que es capaz de aglutinar numerosos morfemas para construir palabras complejas. Por otro lado, el español puede ser clasificado como una lengua fusional, donde las palabras no contienen muchos morfemas distintos, pues varios morfemas pueden ser fusionados o traslapados en uno solo que codifique diversos significados.

Adicionalmente, el náhuatl es una lengua que posee una gran variación dialectal y que no tiene una normalización ortográfica generalizada. Lo anterior dificulta su procesamiento computacional y la disponibilidad de recursos estandarizados, por ejemplo, diccionarios.

Cuando nos enfrentamos a dificultades como las antes mencionadas, los métodos tradicionales de extracción léxica bilingüe necesitan ser replanteados o adaptados. La presente tesis tiene como objetivo proponer una metodología para realizar extracción léxica bilingüe bajo estas condiciones experimentales que hemos descrito de manera general.

Esperamos que la metodología pueda resultar de utilidad para ser aplicada a otros pares de lenguas que enfrenten condiciones similares, especialmente en países como México que poseen una enorme diversidad lingüística pero pocas o nulas tecnologías del lenguaje desarroladas. En el caso del español-náhuatl no existe aún un traductor automático a gran escala, la tarea de extracción léxica bilingüe puede constituir una contribución para alcanzar esa meta en el futuro.

Asimismo, este trabajo puede contribuir en la creación de recursos digitales para este par de lenguas, en particular, corpus paralelos digitales y lexicones bilingües obtenidos de manera automática. Así como profundizar en el tratamiento computacional de lenguas de bajos recursos digitales.

1.2. Objetivos

1.2.1. Objetivo general

Contribuir al desarrollo del procesamiento del lenguaje natural enfocado a lenguas mexicanas de escasos recursos digitales. Particularmente contribuir en la tarea de extracción léxica bilingüe para el par de lenguas español-náhuatl y profundizar así en el estudio de estas lenguas en términos de modelos computacionales estadísticos.

1.2.2. Objetivos específicos

- Construir un corpus paralelo español-náhuatl que permita realizar la experimentación de esta tesis y que, además, sea fácilmente accesible de tal manera que pueda ser aprovechado para el desarrollo de otras tecnologías del lenguaje. Esto implica la búsqueda y recolección de bibliografía bilingüe de fuentes digitales, así como la digitalización, reconocimiento óptico y revisión de textos provenientes de fuentes físicas.

- Proponer una metodología para la extracción léxica bilingüe automática del español-náhuatl. Esta metodología debe ser adecuada para lidiar con las siguientes condiciones experimentales generales: el par de lenguas es distante, se tiene un corpus paralelo pequeño en términos de las magnitudes utilizadas en los métodos tradicionales. Además, una de las lenguas (náhuatl) carece de recursos y herramientas digitales.

- Explorar qué representaciones del texto y qué tipo de información resultan apropiadas para facilitar la tarea de extracción léxica bilingüe entre el español y el náhuatl. Lo anterior implica elaborar modelos computacionales de diversos niveles lingüísticos de las lenguas tratadas. Contribuyendo no solo a la extracción de pares de traducción sino al conocimiento lingüístico desde una perspectiva cuantitativa.

1.3. Hipótesis

Para abordar la tarea de extracción de léxico bilingüe con las restricciones o condiciones experimentales antes mencionadas, proponemos la siguiente hipótesis general:

- Se pueden estimar correspondencias léxicas bilingües a partir de corpus paralelos pequeños (en este caso del español-náhuatl), en primer lugar, por medio de la construcción de representaciones que tomen en cuenta características tipológicas del par de lenguas. Particularmente que consideren el comportamiento morfológico de las lenguas, de tal manera que estas representaciones ayuden a contrarrestar el efecto negativo ocasionado por un corpus paralelo pequeño de lenguas morfológicamente ricas.

Asimismo, es necesario reformular las aproximaciones estándar de extracción léxica bilingüe, esto con el fin de mejorar su desempeño cuando hay pocos recursos disponibles. Lo anterior se puede lograr mediante la combinación de distintos tipos de información para estimar las correspondencias bilingües, esto es, diversas medidas estadísticas de asociación, características contextuales semánticas y representaciones vectoriales.

1.3.1. Preguntas de investigación

- ¿Analizar y tratar la morfología de las lenguas puede facilitar la tarea de encontrar correspondencias léxicas entre una lengua altamente aglutinante y una fusional, por ejemplo, hacer una alineación a nivel subpalabra o morfemas en vez de alinear palabras completas?

- ¿Las representaciones a nivel morfema pueden facilitar el análisis estadístico del corpus, así como disminuir el problema de dispersión cuando se construyen vectores semánticos de palabras para lenguas morfológicamente ricas y con poco corpus paralelo disponible?

- ¿Qué tipo de información y métodos pueden combinarse para estimar de una mejor manera el lexicón bilingüe español-náhuatl?

- Muchos de los métodos que utilizan información contextual de las palabras para estimar las traducciones, requieren en algún punto de un lexicón semilla que les permita comparar los contextos entre dos lenguas ¿Es posible inducir este lexicón inicial de manera no supervisada y así prescindir de uno precompilado?

  De manera general, ¿hasta qué punto nuestra metodología puede mantenerse no supervisada e independiente de las lenguas tratadas, esto es, prescindir de corpus etiquetados, herramientas del lenguaje, conjunto de reglas específicas de la lengua, etc. ?

1.4. Estructura de la tesis

Este trabajo de tesis inicia con la introducción al área de procesamiento de lenguaje natural. Se abordan los fundamentos conceptuales necesarios para el desarrollo del tema

de extracción léxica bilingüe automática, y se ilustran algunas características lingüísticas del español y el náhuatl (capítulo 2). Posteriormente, en el capítulo 3 contiene un panorama de la evolución de los trabajos relacionados con la extracción léxica bilingüe, poniendo particular interés en analizar la capacidad de estos métodos para lidiar con entornos de bajos recursos digitales.

En el capítulo 4 se aborda la metodología general de esta tesis, esto es, un análisis de los retos involucrados en la extracción bilingüe español-náhuatl, así como el diseño de diferentes estrategias para su solución. Lo anterior incluye la construcción del corpus de experimentación, el procesamiento de los textos y la propuesta de diversos métodos para diferentes etapas de la metodología.

El capítulo 5 contiene la evaluación de los diferentes entornos experimentales planteados en esta tesis. En este capítulo se realiza una discusión de las implicaciones en términos lingüísticos y computacionales de la metodología propuesta así como de las estrategias que resultaron favorables para la tarea de extracción de léxico bilingüe español-náhuatl.

Finalmente, el capítulo 6 contiene las conclusiones de este trabajo, un resumen de las aportaciones obtenidas; así como un análisis de las posibles direcciones de trabajo futuro.

Capítulo 2

Antecedentes

2.1. Procesamiento del lenguaje natural

Dentro de las diferentes ramas de estudio de las ciencias de la computación está la inteligencia artificial (IA). Es difícil concretar una definición unificada de IA, puesto que es un concepto que abarca diversos problemas de investigación, diferentes posturas y que ha evolucionado con el tiempo. Por ejemplo, lo que hace algunas décadas se consideraba IA hoy puede ya no serlo; como los primeros sistemas expertos que consistían en una serie explícita de reglas para poder tomar decisiones, sin embargo, no eran capaces de obtener este conocimiento de forma automática. De manera general, podemos decir que la IA se enfoca en la creación de sistemas computacionales que sean capaces, hasta cierto punto, de percibir su entorno y tomar decisiones con algún fin (Nilsson, 2009).

El área ha puesto gran énfasis en explorar la capacidad de las computadoras de imitar capacidades humanas como el aprendizaje, la resolución de problemas, la percepción sensorial, el lenguaje, entre otras. En particular, el interés por tratar el lenguaje humano mediante un sistema computacional dio inicio al área del procesamiento del lenguaje natural.

El PLN comprende modelos, métodos y sistemas computacionales que se especializan en analizar, producir y/o modificar textos (o habla). De tal manera que el texto ya no es visto como una mera secuencia de cadenas alfanuméricas, sino que se procesa con algún conocimiento del lenguaje humano, puede ser muy superficial o profundo, dependiendo de la aplicación. Procesar o modelar el lenguaje humano desde una perspectiva computacional, representa un gran reto, por lo que usualmente es abordado desde un enfoque

interdisciplinario en donde conviven áreas como la lingüística, la computación, las matemáticas, las ciencias cognitivas, entre otras.

La meta de llegar a un total "entendimiento" o imitación del lenguaje humano mediante computadoras quizá aún esté lejana. Sin embargo, hoy en día las tecnologías del lenguaje son una realidad con la que interactuamos cotidianamente. Son ejemplo de esto los asistentes de voz, los buscadores, los sistemas pregunta-respuesta, los traductores automáticos, solo por mencionar algunos.

Podemos rastrear los inicios del área en el año 1949, cuando Warren Weaver sugirió que la teoría matemática de los sistemas de comunicación podría ser aplicada a la traducción automática (Mitkov, 2005). A partir de ahí, el área ha evolucionado dramáticamente, desde los modelos formales inspirados por el conocimiento lingüístico, hasta el PLN de tipo estadístico. En un inicio, la idea era aprovechar el poder de la computadora para ejecutar, con gran velocidad, un conjunto de reglas explícitas escritas por lingüistas. Se popularizó el uso de gramáticas como sistemas deductivos con el potencial de analizar texto automáticamente (Lees y Chomsky, 1957).

En las últimas décadas, el interés se ha centrado fuertemente en enfoques de tipo estadístico que abarcan el uso de inferencia estadística, modelos probabilísticos, aprendizaje de máquina, entre muchas otras técnicas para analizar el texto y generar tecnologías del lenguaje. Un ejemplo de este cambio es la traducción automática, que en sus inicios fue abordada por enfoques formales que requerían un profundo conocimiento lingüístico expresado en reglas, hasta que el reporte ALPAC (American Language Processing Advisory Committee) concluyó que no se había logrado un avance significativo (Pierce y Carroll, 1966). Varios años después, el tema volvió a cobrar relevancia, pero esta vez por medio de enfoques estadísticos que caracterizaban la relación entre traducciones usando corpus paralelos.

Finalmente, cabe decir que el PLN se ha diversificado e interactúa con un número creciente de áreas y problemas de investigación. Por ejemplo, cómo representar a los textos para su procesamiento computacional, el diseño de sistemas de recuperación de la informa-

ción, las interfaces humano-máquina, la gestión inteligente de documentos, la traducción automática, la generación del lenguaje y los modelos computacionales que profundizan el estudio de fenómenos lingüísticos, entre muchos otros (Gelbukh et al., 2006).

2.2. El corpus en PLN

En la disciplina lingüística, un corpus es un conjunto de datos lingüísticos, orales o escritos, que son representativos del uso de la lengua y que se encuentran sistematizados de acuerdo a diferentes criterios de diseño. Se busca que el corpus sea lo suficientemente representativo de la variedad u objeto lingüístico que se pretende analizar (Sánchez, 2001). La existencia de este tipo de recursos ha permitido el estudio de muchos fenómenos del lenguaje humano (lingüística de corpus), constituyen también la base para el desarrollo de diversas tecnologías del lenguaje. Por ejemplo, los corpus orales son fundamentales para construir sistemas de reconocimiento de voz, mientras que los corpus textuales digitales son, actualmente, la base para construir numerosas aplicaciones de PLN (etiquetadores automáticos, clasificadores de documentos, traductores automáticos, entre muchos otros).

Cuando se construye un corpus se deben tener en cuenta ciertos parámetros de diseño, como el dominio al que pertenecen los textos, representatividad, equilibrio, variedad, modalidad del lenguaje usado, entre otros (Sierra Martínez, 2017). Hoy en día en el área de PLN, los corpus suelen ser grandes colecciones digitales de documentos que son procesados fácilmente por computadora.

2.2.1. Corpus paralelos

Existe un tipo de corpus particular al que se le conoce como corpus paralelo, tiene la característica de que todos los documentos que lo forman poseen su respectiva traducción en una o más lenguas. En principio, para considerar que dos documentos son paralelos deben cubrir los mismos significados y tener idénticas funciones en ambas lenguas (Magaz, 2003). Esto los convierte en una gran fuente de información léxica bilingüe que permite el desarrollo de diversas tecnologías del lenguaje, principalmente relacionadas con la traducción.

Algunas de las fuentes tradicionalmente aprovechadas para obtener texto paralelo son manuales técnicos o documentos legales publicados por organizaciones internacionales en diversas lenguas, por ejemplo, el Parlamento Europeo, la ONU, etc. La explosión de la web facilitó tanto el acceso como la generación de grandes cantidades de información; hoy en día la web representa una fuente común para extraer corpus lingüísticos, incluidos los corpus paralelos: "the web as a parallel corpus'' (Resnik y Smith, 2003). La disponibilidad de grandes cantidades de corpus paralelos permitieron el nacimiento de la traducción automática estadística, cuya idea se basa en modelar las relaciones que existen entre dos lenguas, a partir de ver muchos ejemplos de traducciones.

Además de ser la materia prima esencial para la traducción automática, los corpus paralelos permiten el desarrollo de diversas aplicaciones del PLN, por ejemplo, son un recurso útil para la extracción léxica automática, permitiendo la construcción de lexicones bilingües, gramáticas paralelas. Son también usados en los sistemas multilingües de recuperación de la información, en aplicaciones de desambiguación de palabras, entre otras (Widdows et al., 2002; Gómez Guinovart, 2012).

Los textos paralelos también resultan útiles para apoyar la labor de los traductores humanos, pues les permiten buscar palabras o expresiones en estos textos multilingües y observar cómo otras personas tradujeron la misma expresión y en qué contextos es adecuada una traducción (Volk et al., 2014). Asimismo, los corpus paralelos son útiles para realizar análisis lingüísticos contrastivos entre dos lenguas, así como para estudiar cuestiones de adquisición de segunda lengua y otros fenómenos multilingües (Johansson, 2007).

Un aspecto importante que permite explotar la información bilingüe contenida en un corpus paralelo es la alineación. La alineación es el proceso de parear correspondencias bilingües a un nivel específico de correspondencia, por ejemplo, a nivel documento (Braschler y Scäuble, 1998), a nivel párrafos (Gelbukh y Sidorov, 2006), a nivel oración (Brown et al., 1991; Gale y Church, 1993) y finalmente, el nivel más granular y difícil de realizar, la alineación a nivel palabra (tema que profundizaremos en la siguiente sección).

En general, los corpus paralelos utilizados en tareas de PLN se encuentran alineados

a nivel oración. Las técnicas de alineación a nivel oración abarcan métodos superficiales que toman en cuenta la longitud de las oraciones medida en palabras o caracteres, así como técnicas más sofisticadas que involucran restricciones de tipo léxico, cognados, análisis gramatical, etc.

Algunos pares de lenguas, por ejemplo el inglés-chino o inglés-francés, tienen disponibles corpus paralelos de gran tamaño que son extraídos de dominios religiosos, legales, técnicos, etc. Sin embargo, es importante recordar que muchos otros pares de lenguas poseen una cantidad limitada de textos digitales tanto monolingües como bilingües. A este tipo de lenguas se les conoce en PLN como lenguas de bajos recursos digitales y, como ya hemos mencionado, la generación de recursos y su tratamiento computacional representan un reto dentro del área.

Es importante mencionar que, en PLN, además de los corpus paralelos existe un tipo de corpus llamado comparable. Los corpus comparables están formados por documentos entre dos o más lenguas que pertenecen al mismo dominio, sin embargo, no representan propiamente traducciones. Al igual que los corpus paralelos, este tipo de recursos puede ser utilizado en tareas de extracción léxica multilingüe; la Wikipedia en diversas lenguas es un ejemplo de un corpus de tipo comparable.

2.3. Fundamentos de la extracción léxica bilingüe y traducción automática

En PLN, se considera extracción léxica bilingüe a la tarea de obtener automáticamente una lista de pares de palabras que son traducciones (Haghighi et al., 2008). Esto se realiza generalmente a partir de un corpus paralelo, aunque otros tipos de corpus también pueden ser utilizados.

Como mencionamos en la sección anterior, lograr parear un corpus paralelo a nivel palabra no es una tarea trivial. En la figura 2.1 se muestran dos oraciones paralelas checo-inglés, con enlaces o alineaciones que indican la correspondencias a nivel palabra. En el ejemplo se puede notar que algunas palabras en la lengua destino (inglés) pueden no tener nin-

guna relación con las palabras de la lengua fuente (checo); también hay casos en donde una palabra puede corresponder a dos o más en la otra lengua, además de que el orden sintáctico no es necesariamente el mismo en las dos lenguas. Este y otros factores inherentes a la diversidad de las lenguas, hacen que la alineación a nivel palabra sea un proceso considerablemente más complejo que la alineación a nivel oración. Incluso, no es una tarea trivial para los traductores humanos a los que se les pide hacer alineaciones manualmente (Ahrenberg et al., 2000; Melamed, 1995) y se enfrentan a casos problemáticos como expresiones idiomáticas, términos complejos, expresiones multipalabra, etc.

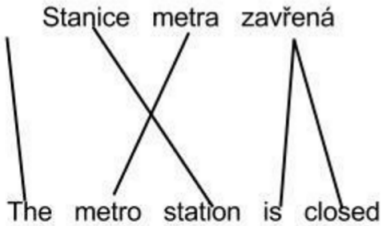

Figura 2.1: Oraciones paralelas checo-inglés, alineación a nivel palabra

Estimar las correspondencias bilingües a nivel palabra a partir de un corpus paralelo es un paso esencial que incorporaron los primeros modelos estadísticos de traducción automática. Esta aproximación a la traducción automática tuvo sus inicios en la década de los 90s y cobró total popularidad en el área alrededor del año 2000. Lo anterior significó un cambió drástico de enfoque, pues durante muchos años el área se había concentrado en diseñar reglas para transformar de una lengua a otra, así como en la obtención de representaciones abstractas del significado (interlingua) que fueran de utilidad para traducir. Como ya hemos mencionado, se considera a la traducción automática como el área que dio inicio a la lingüística computacional, se ha invertido gran cantidad de trabajo a lo largo de varias décadas.

La traducción automática estadística (SMT por sus siglas en inglés) es un enfoque basado en datos (data-driven) que a partir de grandes cantidades de traducciones, corpus paralelos, estima un modelo probabilístico de traducción. La idea general está inspirada en el modelo del canal ruidoso de la teoría de la información (Shannon, 2001) que plantea que un mensaje que se transmite a través de un canal ruidoso puede ser reconstruido por el receptor utilizando el modelo probabilístico de la fuente y del canal ruidoso. De la

misma manera, al hablar de traducción asumimos que una lengua, por ejemplo inglés, es transmitida a través de un canal ruidoso y el resultado es un mensaje en francés. Por lo tanto, el proceso de traducción consiste en reconstruir o decodificar el mensaje original en inglés a partir del mensaje en francés.

La figura 2.2 muestra el modelo del canal ruidoso aplicado a la traducción. Queremos encontrar la mejor traducción e para una oración de entrada f.Para calcular $p(e|f)$ necesitamos un modelo del lenguaje de la lengua fuente $p(e)$ que permita estimar cuál es la probabilidad de una oración en esa lengua, esta distribución se estima a partir de corpus monolingüe. Se necesita también un modelo de traducción $p(f|e)$ que permita estimar la probabilidad de traducción entre oraciones de la lengua destino y la lengua fuente. Esta distribución probabilística se estima a partir de un corpus paralelo, sin embargo, no se modela directamente para oraciones completas; esto resultaría difícil pues la mayor parte de las oraciones solo ocurren una vez, aún en corpus grandes. En su lugar, se realiza un proceso generativo en donde se modelan las probabilidades de traducción a nivel palabra y a partir de esto se estima la traducción a nivel oración (Koehn, 2009).

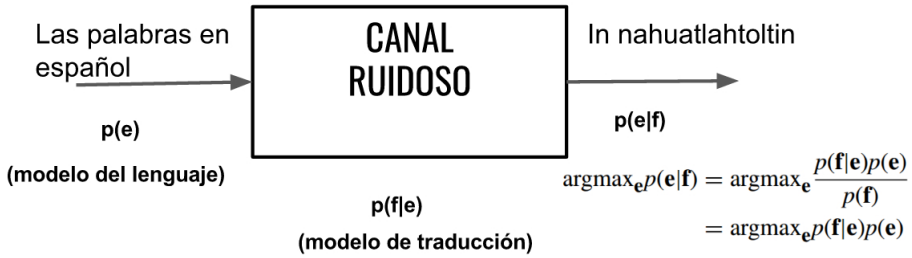

Figura 2.2: Modelo del canal ruidoso utilizado en la traducción automática estadística

Las distribuciones de probabilidad de traducción a nivel palabra, se aprenden a partir de oraciones paralelas y utilizando un algoritmo no supervisado de esperanza maximización. La tarea es vista como un problema en donde existen variables ocultas, las alineaciones que relacionan a las palabras de una lengua con la otra, y se debe estimar el modelo que produce estas alineaciones. El algoritmo se inicializa asignando una distribución uniforme a las posibles alineaciones entre palabras (cada palabra de la lengua fuente puede traducirse como cualquier palabra de la lengua destino con la misma probabilidad). Posteriormente se mejoran estas alineaciones haciendo conteos en el corpus y se construye un

modelo general de estas alineaciones, que va mejorando en cada iteración hasta converger.

El resultado de este proceso son tablas de traducción léxica como se muestra en el ejemplo de la figura 2.3. A partir de estas distribuciones se puede extraer léxico bilingüe o realizar traducción automática de oraciones. Estas nociones fueron planteadas en el modelo IBM-1 que dio inicio a la traducción automática estadística (Brown et al., 1993). Posteriormente se crearon modelos IBM de mayor complejidad, en donde la unidad mínima de traducción no es una palabra sino frases, además de incorporar restricciones en el orden de las palabras, reordenamiento, inserción de elementos nulos para alinear palabras que no tienen equivalente en la otra lengua, entre otras características (Koehn et al., 2003). Sin embargo, la alineación a nivel palabra sigue siendo un paso necesario para este tipo de modelos.

Tengo		**flores**		**azules**	
w	p(w|tengo)	w	p(w|flores)	w	p(w|azules)
nicpia	0.751	xochimeh	0.627	texohqueh	0.235
niquinpia	0.393	xochitl	0.590	yeloh	0.188
onicpiaya	0.265	cuiccaxochitl	0.374	texohtiqueh	0.183

Figura 2.3: Ejemplo, tablas de traducción léxica obtenidas con modelo IBM-1 para español-náhuatl

Es importante mencionar que los modelos utilizados en la traducción automática estadística, se generan a partir de corpus a gran escala. Esto provoca que la calidad de estos sistemas sea dependiente de la cantidad de datos (Germann et al., 2001; Oard et al., 2003) y que por tanto los métodos no puedan ser siempre aplicados exitosamente a varios pares de lenguas, especialmente las que cuentan con bajos recursos digitales. Además, entre más distantes son las lenguas, se necesita mayor corpus de entrenamiento. La tabla 2.1 muestra el tamaño de corpus que se ha utilizado para entrenar sistemas de traducción automática para diversos pares de lenguas (Koehn, 2009).

De manera general, los métodos para encontrar correspondencias bilingües pueden dividirse en enfoques de tipo estimativo y en enfoques de tipo asociativo (Tiedemann, 2003). Los modelos IBM de traducción que se han explicado hasta ahora constituyen el enfo-

Par de lenguas	**Corpus de entrenamiento (palabras)**
francés-inglés	40 millones
árabe-inglés	200 millones
chino-inglés	200 millones

Tabla 2.1: Corpus paralelo utilizado en sistemas de traducción automática

que estimativo, puesto que estiman parámetros o variables ocultas mediante el uso de modelos probabilísticos, así como de un proceso de maximización que produce tablas de traducción léxica. Los métodos de tipo asociativo, por su parte, toman en cuenta medidas de similitud o medidas estadísticas de asociación para determinar la correspondencia entre unidades léxicas. Por ejemplo, coeficiente de Dice (Smadja et al., 1996; Tiedemann, 2000), t-score (Ahrenberg et al., 1998), test de verosimilitud usando log-likelihood (Tufiş y Barbu, 2002), radios de similitud entre cadenas (Melamed, 1995), entre muchos otros.

En el capítulo 3 abordaremos a mayor detalle los trabajos más recientes para realizar extracción léxica bilingüe, utilizando diferentes tipos de métodos.

A partir del año 2014, el área de traducción automática puso gran interés en las arquitecturas de redes neuronales artificiales, dando inicio así a la traducción automática neuronal. El paradigma, en esencia, no ha cambiado, sigue siendo un enfoque estadístico que aprende a partir de un corpus paralelo. La noción en la que se basa este tipo de traducción es que podemos procesar la representación vectorial de una oración en la lengua fuente mediante múltiples capas de una red neuronal y obtener una representación abstracta que codifica características léxicas sintácticas y semánticas. Si aplicamos el mismo procesamiento a la oración paralela en la lengua destino, se esperaría que las representaciones abstractas fueran un tanto similares e independientes de la lengua. Los modelos se entrenan con este conocimiento obtenido a partir de un corpus paralelo para después generalizar y traducir cualquier oración (Alpaydin, 2016; Ruiz Costa-Jussà et al., 2014; Klein et al., 2017).

2.4. Representaciones vectoriales de palabras

En PLN, los textos suelen representarse de manera numérica para facilitar su procesamiento; por ejemplo, muchos métodos necesitan de representaciones vectoriales para poder funcionar adecuadamente. Un documento de texto se puede representar mediante un vector que codifica el patrón de ocurrencias de términos o palabras dentro del documento. Para esto se construye una matriz documento-término que tiene como columnas todo el vocabulario de palabras contenidas en una colección, mientras que los renglones corresponden a los documentos. El valor de cada celda de la matriz está determinado por la frecuencia de aparición de una palabra en ese documento. En PLN, a este tipo de enfoques también se les conoce como modelos de espacios vectoriales.

Estos modelos de espacios vectoriales son también ampliamente usados en tareas de recuperación de la información. La idea general es que al representar cada documento de una colección como un punto en el espacio (un vector), podemos interpretar que aquellos puntos que estén más cercanos, representan a documentos semánticamente similares.

2.4.1. Modelos de semántica distribucional

Además de los documentos, se pueden modelar otras unidades mediante representaciones vectoriales, por ejemplo, palabras, frases, morfemas, etc. Unos de los planteamientos populares para modelar el significado de las palabras mediante vectores, lo constituyen los modelos de semántica distribucional (DSM). En este tipo de modelos, una palabra es representada por un vector que codifica el patrón de co-ocurrencias de esa palabra con muchas otras en un corpus monolingüe de gran tamaño. Este tipo de modelos están inspirados en la hipótesis distribucional establecida en la lingüística, la idea aplicada en PLN consiste en que las palabras que ocurren en contextos similares, tienden a tener significados similares (Wittgenstein, 2010; Harris, 1954; Weaver, 1955; Firth, 1957; Deerwester et al., 1990).

Un DSM puede ser definido por una matriz de co-ocurrencias M, de tal manera que cada fila x representa la distribución de una palabra, o término objetivo, a través de los contextos, las columnas son el conjunto de elementos que representan a los contextos y

que sirven para determinar la similitud contextual entre palabras. Un término objetivo puede ser una palabra, un lema, una frase, un morfema, etc. Mientras que los contextos se obtienen de las ventanas de palabras que rodean al término objetivo (Lund y Burgess, 1996). Los contextos también pueden ser párrafos completos o documentos (Griffiths et al., 2007), información gramatical de las palabras que rodean al término objetivo (Lin, 1998; Padó y Lapata, 2007; Mirza y Bernardi, 2013) y algunos otros contextos enriquecidos.

A la matriz M de co-ocurrencias, de dimensiones $m \times n$, usualmente se le conoce como espacio semántico o matriz de contextos. Esta matriz contiene la información de las co-ocurrencias, cada fila corresponde al vector de una palabra o término objetivo y cada columna a algún contexto lingüístico. La idea es que vectores similares indican significado similar entre palabras. Esta cercanía puede ser calculada usando alguna métrica o medida de similitud como la distancia Euclideana o la similitud coseno.

En general hay varios parámetros que se deben tomar en cuenta cuando se construye un DSM (Baroni y Lenci, 2010). Entre los principales parámetros se encuentran: las unidades lingüísticas que representarán las filas y columnas de la matriz, la longitud máxima de la ventana de palabras que se usa para calcular una co-ocurrencia, el esquema para dar peso (ponderar) a las co-ocurrencias, este puede ser simplemente la frecuencia u otros esquemas como TF-IDF (Sparck Jones, 1972), o medidas estadísticas de asociación como la información mutua (Church y Hanks, 1990). También se debe establecer una medida de similitud para aplicarla a los vectores resultantes.

Finalmente, como los espacios vectoriales resultantes suelen ser de una dimensionalidad muy alta y dispersos, se utilizan técnicas de reducción de la dimensionalidad que comprimen la matriz, tratando de no perder información. Se puede utilizar descomposición en valores singulares (SVD) (Déjean et al., 2002), aunque también se pueden aplicar técnicas de selección de características para solo conservar los contextos más relevantes o informativos.

Los DSM se han vuelto populares en los últimos años por su capacidad de extraer cono-

cimiento automáticamente a partir de un corpus sin necesidad de utilizar otros métodos más costosos para capturar la semántica, como ontologías o bases de conocimiento hechas a mano (Turney y Pantel, 2010). Las representaciones obtenidas con DSM permiten detectar fácilmente diferentes tipos de similitud semántica, por ejemplo, relaciones como la sinonimia, la hiperonimia, entre otras. Debido a lo anterior, numerosas aplicaciones de PLN, hoy en día, parten de representaciones vectoriales de las palabras que se basan en nociones distribucionales de la lengua.

2.4.2. Modelos semánticos distribuidos (*word embeddings*)

En los años más recientes, ha existido un creciente interés en las representaciones vectoriales distribuidas, comúnmente conocidas en la literatura como *word embeddings*. Estos modelos de espacio vectorial también capturan características semánticas de las palabras o unidades lingüísticas y, en esencia, siguen basándose en la hipótesis distribucional como los DSM. Sin embargo, los *word embeddings* son representaciones densas y de menor dimensionalidad que las representaciones semánticas distribucionales.

La idea de estas representaciones distribuidas es combinar espacios semánticos vectoriales con la predicción de modelos probabilísticos. Particularmente, se utilizan modelos neuronales probabilísticos del lenguaje. En PLN, los modelos del lenguaje sirven para estimar cuál es la probabilidad de una secuencia (típicamente palabras) en una lengua, esta distribución de probabilidad se obtiene haciendo conteos de n-gramas en un corpus monolingüe de gran tamaño. A continuación se muestra un ejemplo de cómo se aproxima la probabilidad de una secuencia de palabras, utilizando trigramas. Los símbolos son marcadores de inicio y final de la oración:

$$
\begin{aligned}
& P(Yo, leo, la, tesis) \\
\approx\ & P(Yo|<s>,<s>)P(leo|<s>,Yo)P(la|Yo,leo)P(tesis|leo,la)P(<\backslash s>|la,tesis)
\end{aligned}
$$

Los modelos neuronales del lenguaje, utilizan una arquitectura de red neuronal profunda para estimar la probabilidad de una palabra dado un contexto $P(w|c)$, estas distribuciones se estiman mediante el proceso de aprendizaje de la red neuronal, de tal manera que su capa de salida contiene las probabilidades de cada palabra del vocabulario dado un contexto específico (Bengio et al., 2003). No solo se obtienen las distribuciones de

probabilidad, sino que la capa de salida contiene representaciones vectoriales para cada palabra. Estas representaciones, que se inducen iterativamente a través del proceso de aprendizaje de la red, son capaces de capturar propiedades semánticas y sintácticas de las palabras. Esto se debe a que la sintaxis y la semántica del contexto constituyen características predictivas de las posibles palabras que co-ocurren en el mismo contexto.

Mikolov et al. (2013c) retomaron los modelos neuronales, pero más allá de utilizar las distribuciones de probabilidad, se enfocaron en las representaciones vectoriales de palabras que se pueden extraer de este tipo de modelos del lenguaje. Esto dio inicio al modelo *Word2Vec* (W2V) que hoy en día es una de las representaciones vectoriales más utilizadas. El planteamiento de W2V se basa en los modelos neuronales del lenguaje, sin embargo, realiza varias simplificaciones del modelo para facilitar su entrenamiento en corpus de gran tamaño. Entre las simplificaciones principales se encuentran la reducción del número de capas de la red neuronal, de hecho W2V no es una arquitectura neuronal profunda. También utiliza una técnica conocida como *negative-sampling* para relajar el cálculo del gradiente de la función *softmax*, de manera que para cada palabra solo se tomen en cuenta algunos contextos, en vez de todas las palabras del vocabulario.

Se puede elegir entre dos arquitecturas neuronales distintas al generar representaciones distribuidas con W2V. Por un lado está CBOW, en donde las representaciones distribuidas de las palabras contexto se combinan para predecir la representación de la palabra que está en medio de ese contexto. Por otro lado está Skip-gram, en donde la representación distribuida de la palabra objetivo se utiliza para predecir las representaciones de las palabras contexto (Mikolov et al., 2013a). La figura 2.4 muestra una representación gráfica conceptual de la arquitectura neuronal de estos modelos. Adicionalmente, existen otros modelos para generar word embeddings, como es el caso de Glove (Pennington et al., 2014).

Al igual que los DSM, en las representaciones distribuidas se debe tener en cuenta varios parámetros, principalmente la ventana de palabras del contexto, el tipo de modelo para aprender las representaciones, el número de dimensiones de los vectores de salida y una métrica o medida de similitud para comparar los vectores.

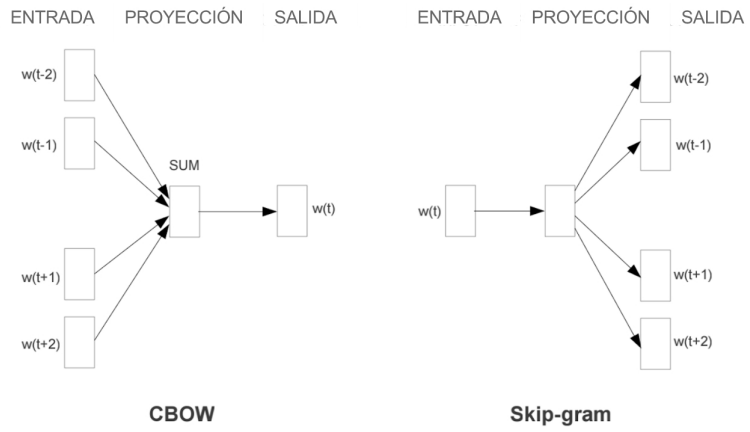

Figura 2.4: Representación gráfica de modelos CBOW y Skip-gram (Mikolov et al., 2013b)

Este tipo de representaciones han cobrado especial relevancia en NLP: capturan fácilmente relaciones semánticas, además, estos tipos de vectores exhiben interesantes propiedades lineales que permiten extraer analogías de palabras, por medio de operaciones vectoriales, por ejemplo, $w_{Paris} - w_{Francia} + w_{Italia} \cong w_{Roma}$.

Finalmente, en los últimos años ha existido un creciente interés por las representaciones distribuidas multilingües, también conocidas como *multilingual word embeddings* (Klementiev et al., 2012; Zou et al., 2013; Hermann y Blunsom, 2013; Irvine, 2013; Lauly et al., 2014; Hermann y Blunsom, 2014). Estas representaciones son útiles para tareas como la extracción léxica bilingüe o mejorar a los sistemas de traducción automática, las discutiremos más a fondo en el siguiente capítulo.

2.5. Características morfológicas del español y el náhuatl

En este trabajo nos enfocamos en el par español-náhuatl, estas son dos lenguas habladas en el mismo país, México, pero pertenecientes a diferentes familias lingüísticas: indoeuropea y yuto-nahua. El náhuatl es una lengua originaria o indígena que posee alrededor de 1.5 millones de hablantes en el país[1] y enfrenta escasez de corpus tanto monolingües como paralelos. Por su parte, el español es una de las lenguas más habladas del mundo y con gran producción de recursos digitales.

[1] http://www.beta.inegi.org.mx/

Al ser de familias lingüísticas diferentes, estas lenguas exhiben fenómenos distintos, por ejemplo, de tipo fonológico, morfológico, sintáctico, semántico. Particularmente, analizaremos el caso de la morfología, es decir, los mecanismos que posee una lengua para formar sus palabras a través de la combinación de morfemas (Haspelmath y Sims, 2010). El náhuatl y el español son dos lenguas morfológicamente ricas pues presentan gran producción flexiva y derivativa; esto quiere decir que son capaces de producir una gran variedad de formas de palabra distintas.

La morfología del náhuatl permite la composición de bloques nominales o verbales compactos, donde los adjetivos, los adverbios y los complementos se aglutinan con los radicales nominales o verbales (Johansson y León-Portilla, 2010), además de que existen fenómenos de incorporación sustantivo-verbo, reduplicación, entre muchos otros. El sustantivo y el verbo del náhuatl, son los tipos de palabra que exhiben más fenómenos morfológicos (principalmente el verbo); en su estructura podemos encontrar diversas celdas de prefijos y sufijos que cumplen diferentes funciones gramaticales (consultar Apéndice B).

El español es una lengua con tendencia fusional, donde varios rasgos morfosintácticos se pueden manifestar en un solo morfema. Por ejemplo, en la forma conjugada del verbo *aman*, el sufijo flexivo *-n* combina diferentes tipos de información: tiempo, aspecto, modo, persona y número.

Otro contraste importante es que la morfología aglutinante del náhuatl permite que dentro de un verbo o sustantivo se codifiquen una gran diversidad de funciones que el español manifiesta a través de su sintaxis. Todas estas características tipológicas provocan que, en las oraciones paralelas (traducciones), sea común encontrar casos en que una palabra del náhuatl corresponde a varias del español; provocando que la oración en español tenga un mayor número de palabras gráficas que su traducción asociada en náhuatl. La tabla muestra un ejemplo de tres oraciones paralelas español-náhuatl, además de la longitud, se puede observar que las oraciones están escritas usando diferentes normas ortográficas del náhuatl. La última oración está escrita en una norma que utiliza letras como *j, k, s*; mientras que las primeras dos oraciones utilizan una norma donde estos fonemas se

expresan con las letras *h, qu/c, z*. Estas diferencias también están relacionadas con la variante, o dialecto, utilizado del náhuatl.

Estos y otros aspectos se abordarán con mayor detalle durante el desarrollo de esta tesis.

Tinechcaquiznequi (náhuatl)
Me quieres oir (español)
In cihuamizton ipan ahcopechtli ca (náhuatl)
La gata estaba encima de la mesa (español)
Pejke san motlajtlachiliyaj (náhuatl)
Empezaron a mirarse nada mas (español)

Tabla 2.2: Ejemplo de oraciones paralelas español-náhuatl

Como a lo largo de nuestra metodología trabajaremos con unidades subpalabra, es importante mencionar que utilizaremos el término *morfo* en vez de *morfema*. Un morfo es la realización en el texto de una unidad abstracta que es el morfema. El término morfo designa un segmento (subcadena) con valor morfológico; no tiene por qué formar parte de un sistema de alomorfos que coinciden en representar un morfema (Bosque, 1983; Hockett, 1971). Primordialmente usaremos este término para referirnos a los segmentos de palabra (subcadenas) que resulten de nuestros procesos automáticos de segmentación morfológica.

Capítulo 3

Avances en la extracción léxica bilingüe

Esta sección contiene un panorama general de la diversidad de aproximaciones que se han diseñado para abordar el tema de extracción léxica bilingüe así como las condiciones experimentales particulares para las que resultan apropiados estos métodos.

La extracción automática de léxico bilingüe generalmente se realiza a partir de corpus paralelos, pero también existen aproximaciones que explotan corpus comparables e incluso grandes corpus monolingües no relacionados. La idea es que estos corpus faciliten el modelado de las correspondencias léxicas que existen entre textos de diferentes lenguas. Como hemos mencionado antes, la tarea de extracción de léxico bilingüe está relacionada con la de alineación a nivel palabra utilizada en los modelos de traducción automática estadística. De hecho, en general, representan la misma tarea, con la salvedad de que la alineación es vista como un paso intermedio necesario para construir un traductor automático, mientras que el objetivo de la extracción léxica bilingüe puede ser la construcción de un lexicón bilingüe *per se*.

El tema de extracción de léxico bilingüe sigue representando un área de investigación activa, pues existe una gran diversidad de situaciones en donde es necesario proponer métodos alternativos o replantear los ya existentes, dependiendo de la naturaleza de las lenguas y de los textos a los que nos enfrentamos.

Por otra parte, los métodos tradicionales como los de tipo estimativo (sección 2.3), se basan en grandes cantidades de corpus paralelos. No todos los pares de lenguas poseen esta cantidad de recursos listos para usarse, existen diversas alternativas para extraer información multilingüe. Algunos autores utilizan diferentes tipos de características para

estimar las correspondencias léxicas en un corpus paralelo, por ejemplo, medidas estadísticas de asociación entre palabras de las dos lenguas, características posicionales de las palabras en el texto, información lingüística, heurísticas, entre otras. En los siguientes subcapítulos describiremos diversos trabajos en torno a la extracción léxica bilingüe.

3.1. Métodos asociativos basados en diferentes características

Existen diversas características de los corpus paralelos que se pueden explotar para extraer correspondencias bilingües. Fung (2000) extrae lexicones bilingües en casos en donde se tiene un corpus paralelo *ruidoso* (lenguas de familias distantes, sin fronteras claras entre las oraciones). Toma en cuenta la distancia entre dos palabras consecutivas (*arrival distance*) para poder comparar las palabras entre dos lenguas y establecer los candidatos de traducción a pesar de que el corpus paralelo no tenga una alineación clara a nivel oración. Este tipo de distancia toma en cuenta las diferencias posicionales entre dos ocurrencias consecutivas de una palabra en el texto, se esperaría que los pares de traducción tengan un comportamiento posicional similar en el corpus paralelo. Lo anterior se implementa mediante vectores que almacenan las diferencias posicionales entre dos palabras consecutivas, así como *dynamic time warping* , una técnica para comparar vectores de diferente dimensionalidad popular en tareas de análisis de señales y reconocimiento de voz.

Moore (2005) realiza alineación a nivel palabra, también a partir de un corpus paralelo. El autor argumenta que los modelos probabilísticos generativos son complejos para implementar y su entrenamiento es lento. En vez de esto propone una gran variedad de métodos basados en medidas de asociación (principalmente *log-likelihood ratio*), obteniendo resultados equiparables a los modelos probabilísticos. Mide asociación entre palabras para obtener alineaciones uno a uno. También construye clusters de candidatos de traducción, para poder extraer alineaciones de muchas palabras a una. Finalmente, también propone incorporar información posicional de las palabras.

Es común que se proponga la combinación de diferentes características con el fin de esti-

mar las correspondencias léxicas bilingües. En Tufis et al. (2006) los autores realizan alineación a nivel palabra combinando diferentes métodos y realizando un pre-procesamiento de los textos que facilita la extracción bilingüe. El pre-procesamiento incluye delimitación de frases y etiquetado gramatical. Entre los numerosos métodos que combinan los autores, se encuentra *log-likelihood ratio* para determinar si un par de palabras se ha visto en oraciones paralelas más de lo que se esperaría por azar. Asimismo, se utilizan redes semánticas, heurísticas para detectar cognados, similitudes de categorías gramaticales, incluso métodos estimativos mediante el alineador GIZA++[1], con el fin de reducir los errores en la extracción léxica bilingüe.

Dentro del conjunto de métodos asociativos que utilizan corpus paralelos, podemos mencionar el método basado en sub-muestreo, también llamado *anymalign*, que es capaz de encontrar correspondencias léxicas a nivel palabra e incluso multi-palabra (Lardilleux y Lepage, 2009; Lardilleux et al., 2011). Este método se basa en la generación de subcorpus aleatorios a partir del corpus paralelo original, de manera que solo las palabras que aparezcan exactamente en las mismas oraciones paralelas dentro de estos pequeños subcorpus, se van tomando en cuenta para calcular un score de traducción. Este método se explorará más a fondo en el capítulo 4 que abarca la metodología de esta tesis.

Existen características que se pueden explotar para encontrar correspondencias bilingües, incluso a partir de textos que no son paralelos. Diversos trabajos se han enfocado en aprovechar la información contenida en un corpus comparable o en corpus monolingües no relacionados para dos lenguas y poder así prescindir de corpus paralelos a gran escala. Toman en cuenta varias características para inducir un lexicón bilingüe, por ejemplo, similitud ortográfica (Koehn y Knight, 2002), temporal (Schafer y Yarowsky, 2002), medidas estadísticas de asociación, información de tópicos (Mimno et al., 2009) y, especialmente, características contextuales. Cuando los pares de traducción se extraen a partir de corpus que no son paralelos, es común que se le conozca a la tarea como *inducción de lexicón bilingüe*.

Las características contextuales se basan en la hipótesis distribucional (Harris, 1954;

[1] www.statmt.org/moses/giza/GIZA++.html

Firth, 1957) pero adaptada a un entorno multilingüe: una palabra que ocurre en cierto contexto debe tener una traducción que ocurra en un contexto similar en la otra lengua. Se puede sintetizar el procedimiento de la mayor parte de los métodos que utilizan información contextual, de la siguiente manera:

1. Se construye un vector que codifique los contextos de cada unidad léxica para cada lengua.

2. Se proyectan estos vectores a un espacio común, de manera que puedan ser comparables. En general, se traducen algunos de los contextos, con ayuda de un diccionario o lexicón semilla.

3. Se calcula la similitud entre los vectores de las dos lenguas para encontrar los candidatos a traducción.

Existen diversos trabajos que hacen uso de información contextual y que varían en la forma en cómo representan los contextos y cómo comparan los contextos entre lenguas para extraer los candidatos a traducción.

Rapp (1995) es uno de los primeros trabajos en extracción léxica bilingüe a partir de corpus comparables utilizando características contextuales. El autor construye matrices contextuales que codifican de manera simple las co-ocurrencias de las palabras en cada lengua. La idea es que al comparar matrices entre inglés y alemán se pueden obtener pares de traducción. Para poder realizar esta comparación, se asigna un índice numérico a cada contexto, se realizan permutaciones para poder cambiar el orden de las palabras en las matrices, así como operaciones matriciales para poder llegar a representaciones equivalentes en las dos lenguas.

Sin embargo, el costo computacional de las permutaciones y comparaciones resulta prohibitivo. Debido a esto, el autor propone una adaptación al método anterior, utilizando un lexicón semilla, de tal manera que se reduzcan las comparaciones y combinaciones posibles (Rapp, 1999).

Gaussier et al. (2004) también realizan extracción de lexicón bilingüe a partir de un

corpus comparable. Sin embargo, los autores señalan que los lexicones semilla comúnmente no tienen buena cobertura, no resuelven cuestiones de sinonimia y polisemia. Para resolver lo anterior, reemplazan el lexicón semilla que permite la comparación o proyección de vectores contextuales, por un mapeo a un subespacio formado por vectores contextuales que se obtienen a partir de las entradas de diccionarios para las dos lenguas, de manera que en este espacio vectorial los sinónimos son vectores cercanos entre sí. Utilizan técnicas de análisis de correlación canónica (CCA), kernels Fisher y LSA multilingüe.

En Haghighi et al. (2008) se combina un modelo estimativo y asociativo para extraer pares de traducción a partir de corpus monolingües para dos lenguas. Los autores inducen lexicones bilingües a partir de un modelo generativo, combinando características contextuales y ortográficas. Se utiliza un lexicón semilla para poder comparar las características contextuales, así como análisis de correlación canónica *(CCA)* y algoritmo de esperanza-maximización (EM) para hacer inferencias en el modelo. Algunas de las limitaciones de este trabajo son que no puede lidiar con pares de lenguas que no tengan similitud ortográfica, que sean tipológicamente distantes, o que los corpus monolingües utilizados pertenezcan a dominios muy distintos.

Prácticamente todos los métodos basados en información contextual requieren en algún punto un diccionario bilingüe semilla. Esto puede representar un problema del huevo y la gallina: Si tenemos un lexicón bilingüe podemos traducir los contextos y comparar vectores contextuales; sin embargo, solo podemos generar un lexicón bilingüe con estos métodos si somos capaces de traducir los contextos(Koehn y Knight, 2002). Existen algunos trabajos que tratan de prescindir de un lexicón semilla pre-compilado, a pesar de usar características contextuales para estimar las traducciones.

En Diab y Finch (2000), los autores proponen un método para encontrar correspondencias léxicas a partir de un corpus comparable y utilizando un lexicón semilla inducido de manera automática, inicializándolo con solo signos de puntuación. Las correspondencias léxicas son encontradas construyendo y comparando "perfiles distribucionales" de las palabras. Los vectores contextuales se construyen en relación a un conjunto reducido de tokens "periféricos" que incluyen los tokens más frecuentes. Para mapear entre lenguas se

busca que sea conservada una medida intralingüe de asociación entre palabras. Al igual que otros métodos, la noción general es que si un par de palabras está altamente asociado en un par de lenguas, entonces estas palabras deben corresponder a un par de palabras que también estén altamente relacionadas en la otra lengua. La medida de asociación se define utilizando coeficiente de correlación de Spearman. Se utiliza un algoritmo de optimización de gradiente descendiente para hacer la búsqueda óptima de pares de traducción basado en las medidas de asociación.

Koehn y Knight (2002) también eliminan la necesidad de un lexicón semilla pre-compilado. Los autores construyen un lexicón bilingüe de sustantivos a partir de corpus monolingües que no están relacionados. Para lograr lo anterior combinan diversas características, por ejemplo, palabras idénticas entre lenguas, palabras con ortografía similar, características contextuales, frecuencias. Inducen un lexicón semilla automáticamente utilizando las palabras con ortografía similar o idéntica. Utilizan un esquema de ponderación (asignación de pesos) para combinar las diferentes características y estimar el mejor par de traducción.

Otra alternativa que resulta conveniente cuando no se tienen recursos para un par de lenguas, es utilizar una lengua intermediaria como lengua pivote para extraer el lexicón bilingüe.

En Seo y Kim (2013) se utiliza una lengua pivote (inglés) para representar vectores contextuales tanto de la lengua destino como de la lengua fuente. Se necesita un corpus paralelo entre la lengua fuente y la lengua pivote, así como un corpus paralelo entre la lengua destino y la lengua pivote. El procedimiento general consiste en construir vectores contextuales para las palabras de la lengua fuente, sus contextos son traducidos a la lengua pivote implementando el alineador de palabras basado en muestreo o *Anymalign* (Lardilleux y Lepage, 2009; Lardilleux et al., 2011). Posteriormente se realiza el mismo proceso para la lengua destino, de tal manera que se pueda calcular la similitud entre los vectores de la lengua fuente y destino, gracias a que sus contextos fueron traducidos con una misma lengua pivote. Una vez calculada la similitud entre vectores de las dos lenguas se realiza un ranking de los candidatos a traducción.

Existen varios ejemplos más de trabajos que explotan una lengua pivote para realizar extracción de léxico bilingüe o traducción automática y que se basan en una noción similar a la antes mencionada. (Tanaka y Umemura, 1994; Wu y Wang, 2007; Tsunakawa et al., 2008).

3.2. Métodos que toman en cuenta la morfología de las lenguas

Resulta importante mencionar los trabajos que lidian con la escasez de recursos no necesariamente proponiendo un nuevo método de alineación a nivel palabra, más bien tratando con la morfología de las lenguas como medio para obtener representaciones más informativas que deriven en un mejor desempeño de los métodos de extracción léxica bilingüe.

Un ejemplo de esto es Nießen y Ney (2004), cuyo objetivo es mejorar los sistemas de traducción automática en los casos en que existe poco corpus paralelo. La principal aportación es introducir conocimiento morfológico para reducir la cantidad de corpus paralelo necesario para entrenar modelos de traducción, tomando en cuenta las dependencias entre palabras flexionadas para que no sean interpretadas como palabras independientes en el modelo. Los autores son capaces de reducir la cantidad de corpus necesario para entrenar a tan solo un 10 % de su tamaño original sin perder un porcentaje significativo de calidad en la traducción. Se utilizan características morfológicas y sintácticas integradas a un modelo probabilístico de traducción automática estadística. En este trabajo es necesario el conocimiento lingüístico explícito de las lenguas analizadas para poder realizar análisis sintáctico y morfológico.

Cakmak et al. (2012) se enfocan en la tarea de alineación a nivel palabra a partir de un corpus paralelo, específicamente para el caso en que las lenguas pertenecen a familias lingüísticas distintas (inglés-turco). Los autores argumentan que la complejidad de alinear lenguas tipológicamente distantes es considerablemente mayor a la de alinear lenguas relacionadas. Se enfocan en este par de lenguas puesto que el turco es una lengua altamente aglutinante con rica morfología, mientras que el inglés exhibe menor riqueza morfológica.

Se realiza alineación a nivel palabra utilizando como corpus de entrenamiento unidades morfológicas en vez de palabras completas con el fin de evitar que una palabra en turco sea alineada con muchas del inglés y así mejorar el desempeño de la tarea. Las correspondencias a nivel morfo se obtienen con un enfoque estimativo (modelos IBM) mediante la herramienta GIZA++.

Otro trabajo que toma ventaja de las representaciones a nivel morfológico para lidiar con entornos de bajos recursos es el de El-Desoky Mousa et al. (2013). Sin embargo, este trabajo no se enfoca en la extracción léxica bilingüe sino en la creación de un modelo del lenguaje para una lengua de bajos recursos y morfológicamente rica, el árabe de Egipto. Los autores construyen un modelo del lenguaje basado en unidades a nivel morfema en vez de a nivel palabra, combinado con redes neuronales profundas. Lo anterior logra una mayor cobertura léxica del modelo del lenguaje y mejorar una tarea de reconocimiento de voz, a comparación de los modelos tradicionales basados en palabras entrenados con pocos datos.

En general, las aproximaciones recientes en PLN de tipo estadístico han empezado a tomar en cuenta la morfología para construir mejores representaciones del lenguaje. Los métodos populares suelen ignorar la morfología de las lenguas, por ejemplo, asignan una representación o vector distinto a cada forma de palabra, aunque estén relacionadas morfológicamente. Esto representa una limitación, especialmente para las lenguas que tienen un vocabulario muy grande y gran producción morfológica; las aproximaciones populares pueden no capturar adecuadamente esta diversidad.

Existen trabajos que exploran cómo las representaciones vectoriales de palabras pueden mejorarse haciendo una composición de las representaciones de los morfemas que forman a las palabras (Lazaridou et al., 2013; Soricut y Och, 2015). Tanto el estudio de fenómenos morfológicos como la elaboración de modelos que tomen en cuenta unidades a nivel subpalabra, constituyen un interés reciente en diversas tareas de PLN. Por ejemplo, se ha aplicado en la traducción automática y aprendizaje de representaciones vectoriales distribuidas para diferentes lenguas (Botha y Blunsom, 2014; Luong et al., 2013; Bojanowski et al., 2016).

3.3. Uso de representaciones vectoriales distribuidas

Finalmente, en los últimos años ha existido un creciente interés por aplicar las representaciones vectoriales distribucionales y distribuidas a entornos multilingües. En esencia, estos trabajos siguen basándose en características contextuales, como los métodos abordados en las secciones anteriores; sin embargo en este caso el tipo de representación vectorial que codifica a los contextos es principalmente de tipo distribuido, es decir, word embeddings.

En este sentido, se han propuesto métodos que buscan aprender representaciones distribuidas bilingües a partir de corpus paralelos, comparables o incluso monolingües. Un ejemplo de esto es el trabajo de Klementiev et al. (2012) donde se inducen representaciones distribuidas conjuntas para un par de lenguas. Se necesita un corpus paralelo para obtener una alineación a priori de un conjunto de palabras (una especie de lexicón semilla), posteriormente se inducen representaciones distribuidas de más palabras utilizando el corpus monolingüe de cada lengua; sin embargo, los pares de traducción obtenidos del corpus paralelo son forzados a estar cerca en el espacio vectorial, con esta restricción se obtiene la representación conjunta de las demás palabras. Los autores modelan el experimento como un proceso de aprendizaje de máquina multi-tarea, donde cada palabra representa una tarea y la relación entre tareas se deriva de alineaciones de pares de palabras obtenidas a partir de un corpus paralelo.

Un trabajo similar pero que no requiere de alineaciones a nivel palabra a priori es el de Hermann y Blunsom (2013). Se enfocan en aprender representaciones bilingües a partir de corpus paralelos alineados a nivel de oración, sin necesidad de tener corpus alineados a nivel palabra o lexicones semilla. Los autores proponen aprender representaciones semánticas composicionales para poder encontrar las correspondencias bilingües a nivel de frases y no solo a nivel palabra. Parten de la idea de que si se tiene un corpus paralelo a nivel oración, se pueden forzar representaciones que capturen los elementos comunes entre enunciados paralelos de diferentes lenguas, es decir, la semántica. Primero se obtienen las representaciones distribuidas de cada palabra perteneciente a una oración, posteriormente mediante operaciones sencillas se combinan estos vectores para obtener una representa-

ción compuesta de la oración. Una vez obtenidas las representaciones vectoriales para cada oración de cada una de las lenguas, se define una función de error para forzar que las oraciones paralelas sean cercanas y que las representaciones de oraciones no paralelas se alejen en el espacio vectorial. Una vez que se entrenó este modelo mediante un proceso de optimización, se puede generalizar y predecir traducciones.

Otro de los primeros trabajos en utilizar representaciones distribuidas es el de Irvine (2013). La autora propone inducir un lexicón bilingüe que sirva para alimentar a las tablas de traducción utilizadas en los sistemas de traducción automática, y así mejorarlos. Este lexicón bilingüe lo induce combinando diferentes fuentes de información. Utiliza un enfoque de aprendizaje supervisado para aprender a combinar las diferentes características (contextuales, temporales, ortográficas), de manera que el clasificador pueda predecir cuando un par de palabras corresponden a una traducción. Las características contextuales son incorporadas a manera de vectores distribuidos obtenidos de corpus comparables y un lexicón semilla para las dos lenguas.

Otra alternativa muy utilizada, y que retoma la idea de proyectar vectores contextuales de diferentes lenguas a un espacio común, son los enfoques que buscan aprender una matriz de transformación que pueda proyectar los vectores de la lengua fuente al espacio de la lengua destino. El trabajo más representativo de esta aproximación es el de Mikolov et al. (2013b), que se enfoca en encontrar un mapeo lineal entre representaciones distribuidas de las dos lenguas con el fin de encontrar correspondencias bilingües. Los autores parten de la idea de explotar propiedades geométricas en los espacios de vectores distribuidos de dos lenguas; asumen que cuando se tiene corpus comparable de gran tamaño, ciertas regularidades permanecen en los dos espacios vectoriales. De hecho, suponen que estas regularidades son lineales y la tarea de traducción de palabras se convierte en aprender una transformación lineal para relacionar los dos espacios.

Para lograr lo anterior se necesita aprender representaciones distribuidas monolingües para cada una de las lenguas y a partir de un lexicón semilla aprender un mapeo lineal entre ellas. El resultado es que si se quiere traducir una palabra de la lengua fuente, se debe proyectar este vector al espacio de la lengua destino mediante el mapeo aprendido.

Una vez que se obtiene un vector en la lengua destino, la traducción será aquella palabra (de la lengua destino) cuyo vector sea el más similar al vector proyectado. Puesto que este tipo de enfoque es de especial relevancia para nuestra metodología, se abordará más a detalle en el capítulo 4. Actualmente existe una gran diversidad de trabajos que proponen variaciones a este planteamiento esencial de aprender una transformación lineal que relacione vectores distribuidos de dos lenguas, estas variaciones estriban principalmente en la representación de los vectores palabra, el tipo de lexicón semilla utilizado, restricciones a la transformación lineal y en el método para recuperar las traducciones más cercanas en la lengua destino (Dinu et al., 2014; Shigeto et al., 2015; Lazaridou et al., 2015; Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017).

Adicionalmente, existen otro tipo de aproximaciones aún poco exploradas y que tienen un interesante potencial de transferencia de conocimiento semántico a lenguas de bajos recursos digitales. La idea es aprender un modelo entrenado para cierta tarea con lenguas de altos recursos y posteriormente transferir o proyectar este conocimiento a una lengua de escasos recursos. Por ejemplo, este tipo de aproximaciones se han utilizado para proyectar anotación de roles semánticos o etiquetas gramaticales a lenguas de bajos recursos digitales, utilizando un corpus paralelo y las anotaciones provenientes de una lengua con mayores recursos (Aminian et al., 2017; Kozhevnikov y Titov, 2013). Así como para realizar traducción automática en entornos de pocos recursos, la idea es primero entrenar el modelo de traducción en un par de lenguas con altos recursos y posteriormente transferir algunos de los parámetros aprendidos al par de lenguas de bajos recursos para inicializar el entrenamiento de su modelo (Zoph et al., 2016).

La tabla 3.1 muestra un resumen de los trabajos representativos expuestos en este capítulo sobre extracción léxica bilingüe, se puede observar que la mayor parte de estos trabajos previos no son aplicados a un escenario de bajos recursos o no consideran la morfología dentro de su metodología. Sin embargo, es importante mencionar que varios de ellos tienen potencial para aplicarse a ciertas condiciones de bajos recursos, sin embargo, la clasificación de "bajos recursos" que se muestra en la tabla se refiere exclusivamente a si el trabajo utilizó explícitamente corpus de tamaño relativamente pequeño.

Trabajo	Características	Corpus	Bajos recursos	Morfología	Lexicón semilla
(Rapp, 1995)	contextuales orden de palabras	comparable	No	No	No
(Rapp, 1999)	contextuales asociación	monolingües	No	No	Sí
(Fung, 2000)	contextuales posicionales	comparable paralelo	No	No	Sí
(Diab y Finch, 2000)	contextuales asociación	comparable	No	No	Sí (inducido)
(Koehn y Knight, 2002)	ortográficas asociación contextuales	monolingües	No	No	Sí (inducido)
(Tiedemann, 2003)	asociación lingüísticas	paralelo	No	No	Sí
(Gaussier et al., 2004)	asociación contextuales sinonimia	comparable	No	No	Sí
(Nießen y Ney, 2004)	estimativas morfológicas sintácticas	paralelo	Sí	Sí	Sí
(Moore, 2005)	asociación posicionales agrupamiento	paralelo	No	No	Sí
(Tufis et al., 2006)	estimativas asociación wordnet	paralelo	No	No	Sí (opcional)
(Haghighi et al., 2008)	contextuales ortográficas	monolingües	No	No	Sí
(Cakmak et al., 2012)	estimativas morfológicas	paralelo	No	Sí	No
(Klementiev et al., 2012)	contextuales rep. distribuidas	monolingües y paralelo	No	No	Sí
(Irvine, 2013)	contextuales rep. distribuidas ortográficas temporales	monolingües y comparable	Sí	No	Sí
(Mikolov et al., 2013b)	contextuales rep. distribuidas	monolingües	No	No	Sí
(Hermann y Blunsom, 2013)	contextuales rep. distribuidas	paralelo	No	No	No
(Seo y Kim, 2013)	contextuales asociación	paralelos	Sí	No	Sí

Tabla 3.1: Resumen de avances en la extracción léxica bilingüe

Capítulo 4

Extracción léxica español-náhuatl

El objetivo general de este trabajo gira en torno al tratamiento computacional de lenguas mexicanas de bajos recursos digitales, en particular, la extracción léxica bilingüe para el español-náhuatl. Este es un escenario experimental con diversas peculiaridades, primeramente, se trata de lenguas para las cuales no es fácil extraer contenido de fuentes digitales: solo es posible obtener corpus paralelos pequeños, en comparación con los utilizados en los métodos tradicionales de extracción léxica bilingüe y traducción automática. El náhuatl no tiene una presencia en la web o producción equiparable al español, muchos de los textos que pueden ser fácilmente encontrados son traducciones. Debido a esto, decidimos constituir un corpus paralelo para estas dos lenguas que fuera de utilidad para realizar la experimentación de extracción léxica bilingüe. El corpus paralelo español-náhuatl constituye una cuestión fundamental para la realización de esta tesis; en la sección 4.1, se explica la tarea de recopilación del corpus, así como los retos que emergieron en su constitución.

Además de la escasez de recursos, las lenguas tratadas son tipológicamente distantes (diferentes fenómenos fonológicos, morfológicos, sintácticos y semánticos), lo que dificulta el uso de modelos que se benefician de características similares entre lenguas. En particular, este par de lenguas exhibe diferencias a nivel morfológico que se explicarán con más detalle en la sección 4.2 y que resultan importantes para encontrar los pares de traducción.

Finalmente, otro factor que es importante tener en cuenta es la gran variación dialectal, así como la falta de normalización ortográfica del náhuatl. Esto provoca que sea difícil conseguir recursos estandarizados, por ejemplo, un diccionario bilingüe estándar que sirva de semilla para inicializar métodos o como fuente de evaluación de pares de traducción.

Bajo estas condiciones experimentales, las aproximaciones tradicionales pueden no resultar efectivas. Por ejemplo, los métodos de alineación a nivel palabra utilizados en la traducción automática estadística típicamente requieren de un gran corpus paralelo. De hecho, hasta donde sabemos, no existe aún un sistema estadístico de traducción automática para el español-náhuatl. Tampoco es posible depender de una tercera lengua pivote que nos ayude a la extracción léxica bilingüe pues los contenidos bilingües que existen en náhuatl suelen ser para español y en mucho menor medida para otras lenguas. Se podría utilizar un enfoque de transferencia de conocimiento semántico, para transferir un modelo de alineación de palabras entrenado en un par de lenguas con mayores recursos a nuestro caso de estudio. Sin embargo, cuando inició este trabajo de tesis no se planeó trabajar con base a corpus de lenguas con mayores recursos y utilizar este tipo de estrategias aún poco exploradas.

Nuestra propuesta consiste, por un lado, en poner particular interés en la morfología de las lenguas como un factor importante para obtener representaciones adecuadas del texto que ayuden a lidiar con la escasez de datos y facilitar así la extracción de correspondencias léxicas. Por lo tanto, los métodos de extracción léxica bilingüe que utilizamos parten de representaciones del texto que toman en cuenta segmentación morfológica y otros tipos de procesamiento morfológico de las palabras.

Por otro lado, nos interesa aprovechar distintos tipos de información o características de los textos paralelos para estimar las correspondencias léxicas entre las dos lenguas. Nuestra conjetura es que la combinación de ciertas características o aproximaciones utilizadas para extraer léxico bilingüe tanto de corpus paralelo como comparable, pueden resultar apropiadas para enfrentarse a un corpus paralelo pequeño y "ruidoso" perteneciente a un par de lenguas distantes. Nuestra intención es combinar métodos asociativos, distribucionales y estimativos. Como se explicó en el capítulo anterior, muchos métodos que hacen uso de características contextuales o distribucionales, requieren de un lexicón semilla bilingüe que ayude a traducir los vectores formados por contextos o a mapear las representaciones entre dos lenguas. Algunos trabajos prescinden de un lexicón semilla extraído de un diccionario, en vez de esto lo inducen a través de similitud ortográfica o

de cognados entre las lenguas. Sin embargo este tipo de aproximaciones usualmente se aplican en pares de lenguas que están emparentadas hasta cierto grado.

Debido a la naturaleza del par de lenguas que estamos tratando probablemente no es posible depender de este tipo de similitudes para formar un lexicón semilla lo suficientemente robusto para funcionar bien en un entorno de pocos datos. Lo anterior debido a que el náhuatl y el español pertenecen a diferentes familias lingüísticas y no comparten una gran proporción de cognados u orígenes etimológicos de las palabras.

Nuestro trabajo pretende prescindir de conocimiento a priori, específicamente de un diccionario semilla precompilado. Proponemos extraer este lexicón semilla a partir de los textos paralelos de manera no supervisada y suficientemente útil para aprender a proyectar representaciones vectoriales de una lengua a otra.

Finalmente, el procedimiento de la extracción de pares bilingües español-náhuatl se puede sintetizar en las dos etapas, muy generales, que se citan a continuación:

- **Pre-extracción de pares de traducción.** Obtención de pares de traducción para formar un lexicón semilla, combinando métodos estimativos y asociativos
- **Obtención final de pares de traducción.** Obtención final de candidatos a traducción, mediante la construcción de representaciones vectoriales multilingües y una transformación lineal aprendida a partir del lexicón semilla

A lo largo de este capítulo abordaremos la metodología de la presente tesis, incluyendo la constitución del corpus de experimentación, el preprocesamiento morfológico de los textos así como las particularidades de nuestro método de extracción léxica bilingüe español-náhuatl.

4.1. Corpus paralelo español-náhuatl

Ya se ha mencionado que los corpus paralelos constituyen un recurso ampliamente explotado en el área de PLN. Las fuentes más comunes para recopilar grandes cantidades

de recursos paralelos incluyen documentos de dominios especializados, como memorias del Parlamento Europeo y de organismos internacionales, textos religiosos y manuales técnicos. Adicionalmente, la web es una buena fuente para encontrar texto paralelo balanceado y de gran tamaño, existen sitios web multilingües que permiten extraer textos paralelos, digitales y listos para usarse. Algunos ejemplos típicos de sitios que ofrecen su contenido en diversas lenguas son: instituciones internacionales, universidades, servicios turísticos, etc.

Sin embargo, no siempre podemos recurrir a las fuentes típicas para obtener recursos lingüísticos, especialmente cuando se trabaja con lenguas de bajos recursos digitales. Estas lenguas son aquellas que tienen una cantidad limitada de recursos digitales a consecuencia de una baja densidad de hablantes, también puede deberse a cuestiones relacionadas con la brecha digital o motivos de otra índole.

En este trabajo enfrentamos un escenario de bajos recursos pues es difícil obtener grandes cantidades de textos paralelos español-náhuatl usando fuentes tradicionales. De hecho, en nuestro caso, la web no representó una fuente suficiente de documentos paralelos, primordialmente utilizamos fuentes no digitales.

Cuando iniciamos la recopilación de traducciones español-náhuatl, nos dimos cuenta que no era fácil extraer contenido de las fuentes web que comúnmente se utilizan para otros pares de lenguas. Los sitios web gubernamentales, turísticos y de otro tipo en México difícilmente ofrecen su contenido en náhuatl, a pesar de ser la segunda lengua nativa más hablada en México y de gozar de carácter nacional. El náhuatl no tiene una presencia web o producción equiparable a la del español, lo que dificulta la extracción de contenido plurilingüe. Es importante mencionar que sí existen algunos recursos en línea para el náhuatl, recursos como Wikipedia. Sin embargo, descartamos el uso de estos textos porque no constituyen propiamente un corpus paralelo sino un corpus comparable (los artículos no son traducciones exactas entre lenguas). Por otro lado, muchos de los que contribuyen con este recurso no son hablantes nativos y la ortografía puede cambiar significativamente de artículo a artículo. Asimismo, no tomamos en cuenta algunos textos de lenguaje muy poético o traducciones de tipo religioso en esta primera recopilación del

corpus paralelo, pues no estábamos seguros de cuál sería el impacto de este tipo de textos en la tarea de extracción automática de léxico bilingüe

La mayor parte de nuestros textos proviene de fuentes no digitales, libros físicos de gran variedad temática. Se realizó una búsqueda exhaustiva en diversas bibliotecas, tanto de la UNAM como externas, con el fin de encontrar libros con contenido paralelo, esto es, traducciones español-náhuatl. Los textos que forman el corpus presentan variación dialectal y diacrónica, es decir, no todos los libros pertenecen a la misma variante de náhuatl ni son del mismo periodo. El náhuatl es una lengua que tiene muchos dialectos, además, los textos no siempre están escritos bajo la misma norma ortográfica, por lo que puede haber diversas grafías asociadas a una misma palabra. Esto se debe a que, aún hoy en día, no existe un consenso sobre la norma ortográfica que resulta apropiada para escribir el náhuatl.

El apéndice A contiene el listado de fuentes físicas con las que se inició este corpus, así como fuentes que ya estaban en formato digital (textos encontrados en internet o que fueron proporcionados por colaboradores). Es importante notar que nuestro proyecto de corpus paralelo español-náhuatl en línea, del cual hablaremos más adelante (sección 4.1.2), permanece activo y constantemente incrementa sus fuentes.

4.1.1. Digitalización y procesamiento del corpus

Una etapa fundamental en el proceso de construcción del corpus, es la digitalización del material bibliográfico, de manera que su contenido pueda almacenarse y procesarse en una computadora. El procedimiento estándar involucra utilizar un escáner para obtener una imagen digital de las páginas de los libros y posteriormente convertir la imagen a texto. Esto se hace mediante un programa conocido como OCR (por sus siglas en inglés Optical Character recognition) que se encarga de reconocer automáticamente las grafías dentro de una imagen.

Para la digitalización del corpus paralelo español-náhuatl se utilizó el OCR Abby FineReader[1], que es un sistema comercial robusto con un buen desempeño en la precisión

[1] www.abbyy.com

del reconocimiento. Sin embargo, el procedimiento estándar de digitalización de un corpus que funciona bien para las lenguas con mayores recursos necesita adaptarse cuando trabajamos con lenguas como el náhuatl. Durante el proceso de digitalización detectamos varios fenómenos interesantes que representan en realidad retos en la tarea de digitalizar textos paralelos español-náhuatl. A continuación describiremos brevemente algunos de ellos:

Ultracorrección. Los OCR son programas que realizan una detección de patrones gráficos para reconocer automáticamente los caracteres contenidos en una imagen; sin embargo, para mejorar la exactitud de este reconocimiento hacen uso de información adicional dependiente de la lengua, por ejemplo, las secuencias de caracteres o n-gramas que son comunes en alguna lengua específica. Cuando el OCR es utilizado para reconocer lenguas como el náhuatl, el software puede no poseer experiencia o entrenamiento previo sobre la lengua, en nuestro caso, el software determinó que se trataba de otra lengua y aplicó ultracorrecciones o falsas correcciones. Nos dimos cuenta de que para palabras en náhuatl, el OCR no reconocía ciertas secuencias de letras, por lo que tendía a asociarlas a palabras del español y las corregía para ajustarlas a las entradas del diccionario en español. Algunos ejemplos de esto son: *itla >ida, yetl >yeti, ye >yo, chalco >chaleo.*

Dificultad para lidiar con marcas fonéticas y tipografía. El náhuatl es una lengua en donde no hay un consenso absoluto sobre la forma de escribir, ocasionando que algunos de los documentos contengan marcas fonéticas o grafías específicas de esa fuente; el OCR mostró dificultad para reconocer algunas de estas marcas, sustituyéndolas por símbolos incorrectos. Algunos ejemplos: *tĕn >ten, síhuat >sihuat, ömpa >ómpa.*

Otro factor que típicamente repercute negativamente es la tipografía de los documentos, existen algunas tipografías que por su diseño provocan que el OCR se confunda fácilmente en la identificación correcta de la letra. Algunos de los libros recopilados presentaban estas características, pues se trataba de ediciones con muchos años de antigüedad.

Errores por el desgaste de las páginas o malas impresiones. El desgaste de la tinta y la mala calidad de la impresión de los libros suele representar un reto para el reconocimiento automático de caracteres. Algunos de los libros recopilados presentaban

estas características, sobretodo ediciones antiguas.

Dificultad para reconocer textos con más de una lengua en la misma hoja. Un interesante fenómeno que detectamos en el funcionamiento del OCR fue la dificultad al enfrentarse a hojas con texto en más de una lengua, en nuestro caso náhuatl y español. El sistema parecía no realizar una identificación de cada lengua contenida en la hoja escaneada, por lo tanto aplicó a toda la hoja el conocimiento de solo una de las lenguas, provocando errores en el reconocimiento.

Errores al separar el texto de las imágenes. Algunos de los libros recopilados, sobre todo los de tipo didáctico, contienen imágenes. El OCR utilizado tiene la funcionalidad de separar las imágenes del texto. Sin embargo, existieron libros que debido a la edición o los colores utilizados en la impresión provocaron que algunas imágenes fueran interpretadas como grafías.

Los problemas antes relatados hicieron necesaria una revisión y corrección manual posterior a las etapas de digitalización y de reconocimiento automático de caracteres. El proceso de corrección manual consistió en comparar las fuentes originales con el texto reconocido por el OCR, en la medida de lo posible. Se modificaron aquellas palabras en donde había errores y se filtró el contenido que no era útil para fines de este corpus. La experiencia adquirida al detectar los patrones comunes de errores cometidos por el OCR fue de utilidad para enriquecer el conocimiento del OCR y mejorar su precisión de reconocimiento paulatinamente. Este enriquecimiento consistió en extender con palabras en náhuatl y español uno de los diccionarios que utiliza el OCR así, como la creación de un alfabeto con diferentes marcas fonéticas presentes en los textos de náhuatl. Todo lo anterior se hizo con el fin de que el OCR cometiera menos errores conforme se fueran escaneando más documentos.

Los documentos paralelos recopilados pertenecen a diferentes dominios, por ejemplo, historia, literatura, material didáctico, recetas. La figura 4.1 muestra una clasificación aproximada de los géneros a los que pertenecen los documentos del corpus paralelo. Además de los diferentes dominios, el corpus presenta variación ortográfica, dialectal y diacrónica,

como ya se ha mencionado antes.

En la tabla 4.1 se muestra una clasificación muy general de la variante dialectal de los documentos, fue realizada con ayuda de un especialista. Los textos más antiguos de nuestro corpus están escritos en náhuatl clásico, esto es, un tipo de náhuatl y una tradición de escritura que se inició en los siglos XVI y XVII. En un inicio, este dialecto se utilizaba para escribir crónicas, textos eclesiásticos, legales, etc. Por otro lado, englobamos dentro de la clasificación de náhuatl moderno, a diferentes variantes que se hablan hoy en día. En la tabla solo se muestra la distribución dialectal del náhuatl, en el caso de las traducciones en español hay mucho mayor uniformidad ortográfica y dialectal, por lo que no se consideró esencial incluir información al respecto en el análisis de la constitución del corpus.

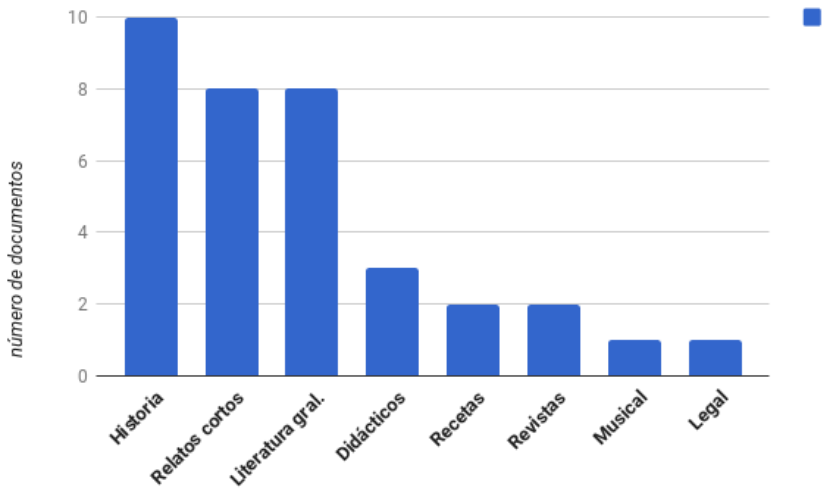

Figura 4.1: Clasificación de los géneros del documento

Dialecto de náhuatl	Porcentaje de documentos
Clásico	45.7 %
Moderno	54.3 %

Tabla 4.1: Clasificación dialectal de documentos en náhuatl

Una vez digitalizados y corregidos los documentos, es importante realizar algún tipo de

alineación o pareamiento entre las dos lenguas, con el fin de aprovechar la información léxica bilingüe contenida en el corpus paralelo. Realizamos alineación a nivel oración, sin embargo, puesto que tratamos con lenguas distantes y con documentos que provienen de fuentes muy variadas, fue necesario utilizar diferentes tipos de técnicas o métodos dependiendo del documento paralelo. Esto es, en algunos casos fue posible utilizar métodos estadísticos tradicionales que se basan en la longitud de oraciones (Gale y Church, 1993; Brown et al., 1991), en otros casos se realizó una alineación semi-automática tomando en cuenta marcadores extraídos del formato de los textos que indican las correspondencias entre fragmentos. En otros casos fue necesario realizar alineación manual con ayuda de humanos que encontraron las correspondencias de oraciones entre textos. Es importante mencionar que debido a la variedad de textos y técnicas empleadas, las alineaciones obtenidas a veces corresponden a unidades más grandes que una oración, por ejemplo, pequeños párrafos. Las oraciones o fragmentos que fueron alineados de manera semi-automática o manual constituyen la mayor proporción en nuestro corpus.

La decisión de alinear gran parte de las oraciones de manera manual o semiautomática fue motivada por una intención de disminuir, en lo posible, los errores de alineación. Los métodos automáticos no siempre tienen una eficacia del 100 %, especialmente si el corpus paralelo es ruidoso o de lenguas distantes, donde no siempre es posible encontrar una alineación uno a uno Singh y Husain (2005). En varios textos aprovechamos el hecho de que poseíamos la imagen digital de la fuente original para observar aquellos marcadores, incluidos en el formato del libro, que indicaran con facilidad la correspondencia entre fragmentos paralelos.

El conjunto de textos que fueron digitalizados, corregidos y pareados a nivel oración o fragmentos constituyen nuestra primera versión de corpus paralelo español-náhuatl. Hasta el momento tenemos 33 fuentes de textos paralelos, el tamaño total del corpus es alrededor de 1,186,662 tokens, tomando en cuenta los documentos de las dos lenguas.

4.1.2. Axolotl, sistema web para consulta del corpus

Una vez construido el corpus paralelo, estábamos interesados en desarrollar una aplicación que permitiera acceder fácilmente a este recurso por cualquier persona. Desarro-

llamos un sistema de recuperación de la información, llamado Axolotl, que mediante una interfaz web permite buscar palabras y frases dentro del corpus en las dos lenguas.

La idea del sistema es realizar búsquedas en los documentos del corpus y desplegar aquellos fragmentos paralelos que contengan el término buscado en español o náhuatl. A través de la interfaz web, los usuarios tienen la posibilidad de hacer búsquedas de palabras o frases en las dos lenguas. Esta recuperación de fragmentos u oraciones paralelas es posible gracias a que el corpus fue alineado a este nivel.

Nuestro sistema web es parecido a otros sistemas de búsqueda en corpus paralelos, por ejemplo, Linguee[2] y OPUS Corpus Query (Tiedemann, 2012). Los resultados de búsqueda desplegados por nuestro sistema Axolotl, no solo contienen los fragmentos textuales, sino información acerca de la fuente de estos fragmentos y una vista previa en PDF que contiene parte del documento original de donde fue extraído.

Adicionalmente, el motor de búsqueda ofrece flexibilidad en el tipo de búsquedas, se pueden realizar búsquedas por palabras o frases, así como utilizar operadores para realizar búsquedas más complejas y obtener resultados más precisos. Por ejemplo, operadores binarios (AND, OR), operadores de proximidad para encontrar frases con palabras a una distancia específica, búsquedas difusas que permiten encontrar términos aproximados al patrón de búsqueda sin necesidad de que coincidan exactamente, etc.

En cuanto a la arquitectura del sistema, su desarrollo incluyó la integración de diversas tecnologías. Por un lado, tratamos que el sistema pudiera manejar eficientemente grandes cantidades de corpus, aunque nuestro corpus paralelo fuera relativamente pequeño. Para esto, se almacenó el texto paralelo, así como las relaciones entre oraciones paralelas, en una base de datos MongoDB[3] que permite escribir y leer los datos de una manera rápida, a comparación de otras soluciones. Con el fin de recuperar eficientemente los resultados de las búsquedas de los usuarios, es necesario tener un motor de búsquedas, en nuestro caso lo implementamos utilizando las tecnologías de Lucene/Solr[4] que son herramientas

[2]http://www.linguee.com/
[3]https://www.mongodb.org/
[4]https://lucene.apache.org/

ampliamente conocidas en el ámbito de la recuperación de la información. El motor de búsqueda es un componente esencial que permite la indexación de los textos paralelos almacenados en la base de datos, para posteriormente recuperar los fragmentos paralelos que responden a la búsqueda de un usuario, así como ordenarlos por nivel de relevancia. Finalmente la interfaz entre estos componentes y el usuario es una implementación web en Ruby on Rails[5]. Se puede consultar información más detallada sobre el diseño y desarrollo de este sistema en Gutierrez-Vasques et al. (2016).

El sistema de búsqueda Axolotl es gratuito y accesible desde la Web [6] (Figura 4.2). Aunque el sistema permite hacer búsquedas en todo el corpus, actualmente no es posible descargar los textos completos del corpus paralelo, puesto que es necesario revisar los derechos de autor aplicables a cada caso antes de publicarlos en algún sitio. En el caso de este corpus paralelo fue necesario revisar la Ley Federal de Derechos de Autor, en particular el derecho patrimonial, que es el derecho del autor para explotar o autorizar a otros explotar sus obras en cualquier forma. Muchos de los textos escritos en náhuatl poseen carácter de dominio público, lo que facilita su difusión a través de medios digitales; sin embargo, son las traducciones al español las que poseen mayores restricciones legales. En resumen, estas restricciones legales previenen en algunos casos la difusión de un libro digitalizado completo, aunque se permite mostrar fragmentos, especialmente si es con fines educativos o de investigación. Debido a esto, por el momento el sistema solo permite la consulta de todos los fragmentos que contienen la palabra o frase buscada.

Hasta donde sabemos, este es el primer corpus paralelo español-náhuatl digital disponible para su consulta en línea. Nuestro objetivo es que este recurso sea de utilidad para impulsar la creación de tecnologías del lenguaje para este par de lenguas, especialmente en un país como México que tiene una vasta diversidad lingüística pero muy pocas tecnologías desarrolladas para las lenguas nacionales.

[5] http://rubyonrails.org/

[6] www.corpus.unam.mx/axolotl

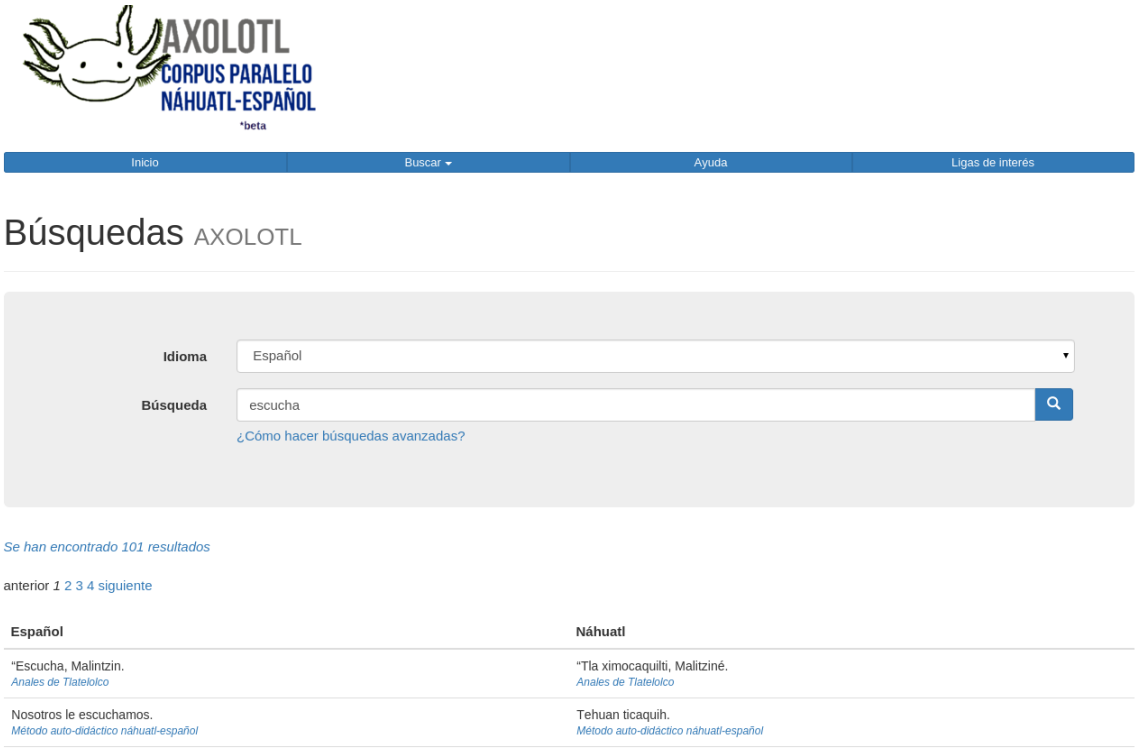

Figura 4.2: Sistema web Axolotl, corpus paralelo español-náhuatl

4.2. Procesamiento morfológico de los textos

4.2.1. Normalización ortográfica

Como mencionamos antes, una vez recopilados los textos fue evidente la falta de normalización ortográfica en la escritura náhuatl. No solo se encuentran grafías muy diferentes entre los textos más tempranos, también los textos actuales presentan gran variación en la escritura. Lo anterior puede ser especialmente contraproducente para los métodos de PLN, tanto para los que se basan en reglas, pues necesitan de la codificación de reglas explícitas para capturar toda esta variabilidad; como para los de tipo estadístico, que al estar basados en frecuencias de las palabras, se benefician considerablemente de una escritura más consistente.

Debido a esto, un paso esencial para poder realizar la extracción léxica bilingüe, es tener textos con cierta normalización ortográfica. Desafortunadamente, en el caso del náhuatl existe una falta de consenso sobre la norma ortográfica adecuada, incluso entre especialistas, entes gubernamentales, etc. El objetivo de este trabajo no es profundizar o proponer

una norma ortográfica, lo que es de nuestro interés es tener consistencia en los textos recopilados para poder aprovecharlos sin tener variaciones que puedan significar "ruido" al utilizar métodos de procesamiento del lenguaje natural.

Decidimos explorar qué alternativas existían para realizar normalización ortográfica automática del náhuatl. Al ser una lengua de bajos recursos digitales, es difícil encontrar herramientas dedicadas al procesamiento del náhuatl, sin embargo, retomamos el trabajo del investigador Marc Thouvenot quién diseñó un conjunto de reglas para normalizar el códice Florentino. La normalización propuesta da como resultado textos con una escritura muy parecida a la que propone Carochi (1645) en su gramática pero sin marcar el saltillo ni la longitud vocálica. Este proceso de normalización se basa en un conjunto de 270 reglas que originalmente estaban estructuradas como se muestra en la tabla 4.2. El ejemplo contiene la cadena que hay que transformar, la nueva cadena, así como la posición o contexto necesario para aplicar la transformación y el orden en que se debe aplicar la regla.

Cadena original	**Cadena norm.**	**Pos.**	**Orden**	**Ejemplo**
as	ax	indif.	2	cacastli>cacaxtli

Tabla 4.2: Ejemplo de regla de normalización ortográfica

Decidimos reimplementar este conjunto de reglas usando el software FOMA, una herramienta que permite crear autómatas y transductores de una manera sencilla y eficiente a partir de reglas, se utiliza principalmente para tareas morfológicas (Hulden, 2009). Sin embargo, es importante destacar que estas reglas fueron creadas originalmente para textos particulares pertenecientes a un periodo, por lo tanto, no son suficientes para normalizar toda la variedad existente en nuestro corpus. Aún así, no quisimos desaprovechar el conocimiento lingüístico codificado en estas reglas para aplicarlas a un subconjunto reducido de nuestro corpus. La normalización de todos los textos en náhuatl requeriría de un trabajo exhaustivo, con ayuda de lingüistas.

En vista de lo anterior, para fines de esta tesis decidimos trabajar solo con un subconjunto de los documentos del corpus paralelo, seleccionamos aquellos para los que fue

posible normalizar o que poseían una escritura más o menos sistemática y similar. Primordialmente son textos escritos en náhuatl clásico con una escritura parecida a la que propone Marc Thouvenot en su trabajo de normalización del Gran Diccionario Náhuatl[7]. La tabla 4.3 muestra el tamaño del corpus paralelo que se utilizó para los experimentos de extracción léxica español-náhuatl.

Hasta el momento no se ha aplicado ninguna normalización ortográfica al corpus en línea Axolotl, por lo que el texto consultable en línea es el original.

Lengua	**Tokens**	**Tipos**	**Oraciones**
Español (ES)	118364	13233	5852
Náhuatl (NA)	81850	21207	5852

Tabla 4.3: Tamaño del corpus paralelo utilizado para experimentación

4.2.2. Segmentación morfológica

El primer paso de nuestra extracción léxica bilingüe, lo constituye el procesamiento morfológico de las lenguas tratadas. Como ya se ha mencionado antes (sección 2.5), el español y el náhuatl son dos lenguas tipológicamente distantes, lo cual implica que no compartan muchas características en términos de morfología, sintaxis, ortografía, etc.

Tomar en cuenta características tipológicas de las lenguas puede resultar de utilidad para mejorar la estimación de correspondencias bilingües. En este sentido no solo es importante poner atención en los métodos de alineación de palabras, sino en la representación morfológica de los textos. Una de nuestras intuiciones iniciales fue que lidiar con la morfología, normalizarla, podría ser de ayuda para reducir el impacto negativo de la escasez de datos y de la dispersión en las representaciones del texto, pues mientras más productivo es el sistema de flexión de una lengua, mayor es el número de formas morfológicamente distintas que pueden ser generadas por esa lengua. Dicho en otras palabras, la riqueza morfológica de las lenguas puede provocar que haya muchos tipos en el corpus, pero pocas repeticiones de estos tipos, lo cual representa un problema si queremos caracterizar una

[7]http://www.gdn.unam.mx

palabra por su frecuencia de aparición o por los contextos en los que aparece.

Adicionalmente, en el caso del náhuatl, por su naturaleza aglutinante con tendencia polisintética, puede resultar de utilidad segmentar los prefijos y sufijos que se aglutinan a los radicales, enfocándose después en los morfemas con contenido léxico y no en los de contenido gramatical para encontrar las correspondencias entre las unidades léxicas de las dos lenguas. También, al desaglutinar las palabras, provocaremos que haya más repeticiones de cada unidad o palabra gráfica en el texto, lo cual facilita la generación de modelos estadísticos. La tabla 4.4 muestra un ejemplo de palabras segmentadas morfológicamente en náhuatl y su traducción al español.

ti - c - cohua - z ("lo comprarás")
2SG.S-3SG.O-'comprar'-FUT
ni - c - cohua ("lo compro")
1SG.S-3SG.O-'comprar'
ni-c-cohua-tica ("lo estoy comprando")
1SG.S-3SG.O-'comprar'-PROG
Correspondencia léxica buscada: *comprar-cohua*

Tabla 4.4: Ejemplo morfología náhuatl

Tradicionalmente, en el área de PLN, la variación morfológica de los textos se normaliza realizando un proceso de stemming o lematización, cuyo objetivo es la eliminación de las terminaciones flexivas de las palabras para obtener raíces o entradas de diccionario (lemas). Esto resulta especialmente útil para lenguas con morfología menos rica como el inglés o para lenguas primordialmente sufijales, como el español (Fábregas, 2013).

En el dominio de extracción léxica bilingüe, la morfología es una cuestión que no se ha explotado extensivamente, al menos no más allá del procedimiento estándar para normalizar el texto utilizando stemming o lematización. Como se vio en la sección 3, existen

trabajos que abordan la morfología más a fondo como medio para mejorar la extracción léxica bilingüe, de tal manera que requieran menos corpus de entrenamiento y puedan tratar con pares de lenguas morfológicamente ricas (Nießen y Ney, 2004; Cakmak et al., 2012). Sin embargo, es importante mencionar que, en este tipo de trabajos, el análisis morfológico suele hacerse mediante herramientas especificas de la lengua.

Sin embargo, en el caso del náhuatl, enfrentamos una escasez de tecnologías para procesamiento morfológico, además de que la morfología del náhuatl requiere de un tratamiento especial, más allá de truncamientos al final de las palabras. En vista de lo anterior, decidimos realizar segmentación morfológica automática.

Hasta donde sabemos la única herramienta computacional disponible para análisis morfológico del náhuatl es Chachalaca (Thouvenot, 2011), un software basado en reglas enfocado en el náhuatl clásico, que es capaz de proponer uno o más análisis morfológicos posibles para una palabra proporcionada. Gracias al contacto con el desarrollador de esta herramienta, el investigador Marc Thouvenot, pudimos analizar su funcionamiento. Se basa en esquemas constructivos que son representativos para la mayor parte de expresiones en los textos del siglo XVI, como se muestra en la tabla 4.5. El programa se basa en las gramáticas de Launey y Thelma Sullivan para establecer listas de morfemas (Launey y Kraft, 1992; Sullivan y León-Portilla, 1976); emplea una lista de 62 prefijos, 173 sufijos y 11,864 formas léxicas (radicales verbales y nominales), además de listas con excepciones y un conjunto considerablemente grande de reglas que definen las combinaciones permitidas[8].

prefijo(s) + r.v. + sufijo(s)
prefijo(s) + r.n. + sufijo(s)
prefijo(s) + r.n.+r.v. + sufijo(s)
prefijo(s) + r.n.+r.n. + sufijo(s)
prefijo(s) + r.n.+r.n.+r.n. + sufijo(s)
prefijo(s) + r.v.+lig.+r.v. + sufijo(s)

Tabla 4.5: Esquemas morfológicos de Chachalaca

[8]También se incluyen ligaduras (lig.), esto es, morfemas que no cargan significado, más bien sólo cumplen una función morfofonológica para unir diferentes morfemas.

A pesar de que esta herramienta representa una gran aportación y posee conocimiento profundo sobre la morfología de la lengua codificado en numerosas reglas, hallamos algunas desventajas que dificultaron su integración a nuestra tarea. Por ejemplo, depende de diccionarios de morfemas, por lo que no puede realizar el análisis de una palabra si un morfema no se encuentra en estas listas, no puede procesar ciertos fenómenos morfológicos del náhuatl, por ejemplo, la reduplicación. Además, es una herramienta de escritorio creada hace varios años y sin actualizaciones, lo cual dificulta su portabilidad a diferentes sistemas operativos, así como su conexión con otros programas. Finalmente, también encontramos que cuando se analizan cantidades relativamente grandes de palabras su rendimiento fue prohibitivo. La actualización de la herramienta Chachalaca, constituye una interesante e importante tarea que puede abordarse como trabajo futuro.

Enfrentamos dificultades al tratar de manipular Chachalaca e integrarlo a nuestra línea de experimentos debido a diversos factores, como el lenguaje de programación en el que está hecho y que solo posee una interfaz de usuario gráfica disponible para Windows. Sin embargo, logramos construir una interfaz de comunicación con Python que no era del todo estable: el tiempo de análisis de documentos grandes era prohibitivo.

Decidimos explorar alternativas no supervisadas, es decir, que no dependen de un conjunto de reglas o información específica de la lengua. Existen aproximaciones de la lingüística computacional que tratan de descubrir automáticamente las fronteras entre morfemas de una palabra, a través del análisis de un corpus usando métodos estadísticos. En particular, existe una herramienta de segmentación morfológica no supervisada que ha gozado de popularidad en el área en los últimos años: Morfessor (Creutz y Lagus, 2005; Virpioja et al., 2013).

Morfessor está diseñado para lidiar predominantemente con lenguas de morfología concatenativa, donde el número de morfemas por palabra puede variar mucho y no se conoce de antemano. Ha probado su buen desempeño en lenguas como el finlandés. Es un algoritmo de tipo no supervisado, por lo que no necesita conocimiento a priori, esto es, un corpus etiquetado con las segmentaciones correctas a partir del cual generar un modelo

de aprendizaje. Sin embargo, permite también utilizarse de manera semisupervisada, con un pequeño conjunto de palabras segmentadas manualmente que sirve para ajustar los parámetros de la segmentación automática.

Hoy en día existen diferentes versiones de Morfessor, sin embargo, en esencia están basadas en modelos probabilísticos generativos que hacen uso del principio de longitud de descripción mínima (MDL). Este principio se inspira en nociones de la teoría de la información, la idea es que cualquier regularidad en los datos puede usarse para comprimir estos datos (describirlos utilizando menos símbolos), de tal manera que entre más se compriman, aprendemos más sobre estos datos. Esto es útil como criterio para elegir un modelo que describa mejor los datos (Grunwald, 2004). Esto no está alejado del estudio lingüístico, pues las lenguas naturales están relacionadas con el concepto fundamental de teoría de la información de incertidumbre cuando se codifica y descifra un mensaje.

En particular, Morfessor ejecuta dos etapas principales en el proceso de segmentación morfológica: el entrenamiento y la decodificación. La fase de entrenamiento recibe como entrada un corpus, que no necesita estar anotado, y a partir de este se calculan los parámetros del modelo de segmentación. Una vez obtenido el modelo, el algoritmo de decodificación se encarga de segmentar nuevas palabras.

En vista de lo anterior, decidimos utilizar Morfessor 2.0 (Virpioja et al., 2013) para la segmentación morfológica de nuestros textos. Trabajamos de manera semi-supervisada, esto es, con corpus monolingües sin ningún tipo de etiquetado (conjunto de entrenamiento) y con un conjunto de desarrollo (*development*) que contiene palabras segmentadas morfológicamente por algún especialista; este conjunto sirve durante el entrenamiento para mejorar las segmentaciones que infiere el modelo. Asimismo, es necesario un conjunto de prueba o *gold-standard*, que no interviene en el entrenamiento, y que contiene también palabras con su segmentación correcta, sirve para evaluar qué morfemas logró segmentar correctamente el modelo aprendido.

En el caso del náhuatl, el corpus de entrenamiento está formado por el conjunto de textos en náhuatl presentados anteriormente en la tabla 4.3, además de algunos textos mono-

lingües que no forman parte del corpus paralelo. En cuanto a los conjuntos de palabras anotados, fueron construidos con la ayuda de especialistas en la lengua, en particular, se ocupó un conjunto de textos segmentados por el investigador Leopoldo Valiñas Coalla.

Decidimos también realizar segmentación morfológica del español, en este caso utilizamos como corpus de entrenamiento los textos en español del corpus paralelo (tabla 4.3), además de un corpus monolingüe de gran tamaño disponible para esta lengua, el Corpus del Español Mexicano Contemporáneo (CEMC) (Lara, 1979). Los conjuntos de desarrollo y prueba son los utilizados en el trabajo de Méndez-Cruz et al. (2016) sobre segmentación morfológica del español. En la tabla 4.6 se muestra el tamaño final de los corpus utilizados para cada lengua en la tarea de segmentación morfológica.

Español	**Tokens**	**Tipos**
Entrenamiento	2175533	99564
Desarrollo	800	800
Prueba	792	792
Náhuatl		
Entrenamiento	83229	22174
Desarrollo	1379	1379
Prueba	288	288

Tabla 4.6: Tamaño de corpus utilizados para generar modelos de segmentación morfológica

Diversos parámetros influyen en la generación de modelos de Morfessor, parámetros como el tipo de entrenamiento para generar el modelo (batch training, online training), el tipo de algoritmo para la etapa de decodificación (viterbi, recursivo) así como la forma en la que son cuantificadas las palabras del corpus de entrenamiento para generar un modelo. Esto es, se pueden tomar en cuenta las frecuencias de aparición de cada palabra de tal manera que esta frecuencia influya en el modelo de segmentación, se pueden también tomar en cuenta sólo a los tipos (palabras gráficas distintas) sin importar su frecuencia de aparición o se puede utilizar el logaritmo de la frecuencia de cada palabra para generar el modelo.

Probamos diferentes parámetros para generar los modelos de segmentación, en todos los casos el entrenamiento fue de tipo batch, y el algoritmo de decodificación, recursivo. En particular, probamos diversos valores del parámetro α (*unannotated corpus likelihood weight*) que controla la sobresegmentación o subsegmentación del modelo (Smit et al., 2014), así como diferentes formas de contabilizar la frecuencia de aparición de una palabra en el corpus de entrenamiento. En la tabla 4.7 se muestran únicamente los modelos que obtuvieron el mejor y peor desempeño para cada lengua (el desempeño detallado para todos los parámetros puede consultarse en el apéndice C).

Se evaluó el desempeño de los modelos con la métrica conocida como Boundary Precision and Recall Evaluation (BPR) (Virpioja et al., 2011), ampliamente utilizada en las tareas de segmentación morfológica. Esta métrica sirve para evaluar qué tan bien se detectaron las fronteras entre morfemas de una palabra, tomando como referencia el gold-standard o conjunto de prueba. BPR consiste en las siguientes medidas de desempeño:

$$Precision = \frac{Num.\,de\,fronteras\,correctamente\,encontradas}{Total\,de\,fronteras\,encontradas}$$

$$Recall = \frac{Num.\,de\,fronteras\,correctamente\,encontradas}{Total\,de\,fronteras\,en\,gold\,standard}$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

La evaluación se realiza de la siguiente manera: se segmenta automáticamente un texto para el cual poseemos las segmentaciones de referencia anotadas por un humano (conjunto de prueba), gracias a lo cual es posible comparar qué fronteras morfológicas detectadas por el programa fueron correctas. La precisión en este caso nos indicaría, de las fronteras encontradas, cuántas fueron de hecho correctas, mientras que el recall indicaría cuántas fronteras fueron correctamente encontradas del total de fronteras que poseía el conjunto de referencia. La medida F-score es una combinación de ambas y es quizá el indicador más informativo, pues el valor de la precisión o el recall por si solo puede ser engañoso, por ejemplo, casos en que la precisión resulte muy alta pero solo debido a que el modelo encontró muy pocas fronteras del total de fronteras que contiene el texto; pero si las pocas

fronteras encontradas fueron acertadas, entonces la precisión sería alta, aunque se haya dejado sin segmentar a la mayor parte de las palabras. Es importante mencionar que en la evaluación consideramos más de una segmentación posible por palabra en el conjunto de prueba.

α	**Precisión**	**Recall**	**F-score**
Español (ES)			
0.4	84 %	80.6 %	**82.3** %
10	98.5 %	17.3 %	29.4 %
Náhuatl (NA)			
0.8	75.1 %	80 %	**77.5** %
10	97.6 %	22.9 %	37.1 %

Tabla 4.7: Evaluación de modelos de segmentación morfológica

A pesar de que la segmentación morfológica obtenida es perfectible, utilizamos estos modelos para segmentar los textos del corpus paralelo. Nuestra intuición fue que aún cuando el segmentador cometiera errores, esto no necesariamente impediría que se normalice la variación morfológica en los textos. Podemos compararlo con el caso del stemming: el hecho de que las cadenas resultantes de este proceso no equivalen necesariamente a una raíz o a un lema de la lengua en cuestión, no significa que no sean informativas y por tanto no impide que el stemming sea un método efectivo para pre-procesar y normalizar morfológicamente los textos.

Por otro lado, conjeturamos que aunque nuestros modelos cometan errores de segmentación, aquellas subcadenas más recurrentes, y por lo tanto más segmentadas, corresponderían a morfemas correctos de la lengua, principalmente con función gramatical. Nos interesa poder filtrar este tipo de morfemas del náhuatl para poder realizar la alineación de correspondencias bilingües. El apéndice D muestra una tabla con los 31 morfos del náhuatl que resultaron más frecuentes al segmentar el corpus paralelo. Resulta interesante notar que todos los morfos más frecuentes corresponden a morfos gramaticales correctos

del náhuatl.

Finalmente, es interesante notar que a pesar de tener muchos menos recursos de entrenamiento disponibles, la segmentación morfológica automática del náhuatl obtuvo resultados equiparables a la del español. Esto podría ser un indicador de que la estructura morfológica aglutinante del náhuatl es relativamente fácil de predecir, a pesar de que sea compleja en términos del número de morfemas que contiene una palabra. Sin embargo, esta hipótesis amerita mayor trabajo futuro.

4.2.3. Relación tipo-token entre textos paralelos

Ya hablamos antes sobre nuestra hipótesis de que lidiar con la morfología nos puede ser útil para reducir el número de formas distintas en el texto y para facilitar el proceso de encontrar correspondencias léxicas si se tiene segmentado morfológicamente el náhuatl. Para explorar de manera más cuantitativa esta noción decidimos calcular la relación entre los tipos y los tokens en los textos paralelos.

La relación tipo-token (TTR) es la relación que existe entre el número de palabras gráficas distintas (tipos) y el total de palabras gráficas (tokens) en un texto. Entendiendo palabra gráfica como una secuencia alfanumérica separada por espacios. Esta medida ha sido utilizada para diversos fines a lo largo de los años, por ejemplo, como un indicador de riqueza léxica y de estilo (Herdan, 1966; Stamatatos, 2009), flujo de información en un texto (Altmann y Altmann, 2008) y también ha sido utilizada en estudios de adquisición del lenguaje, de psiquiatría y de literatura.

Recientemente, TTR ha probado ser una forma simple, aunque efectiva, de cuantificar la complejidad morfológica de una lengua a partir de un corpus (Kettunen, 2014; Bentz et al., 2016). Además, cuando se calcula en un corpus paralelo, TTR es útil para contrastar características tipológicas de las lenguas (Kelih, 2010).

Desde la perspectiva lingüística, existen diversas aproximaciones y definiciones sobre la complejidad morfológica de una lengua (Baerman et al., 2015; Anderson, 2015; Sampson et al., 2009), no es una noción fácil de conceptualizar y cuantificar, pues intervie-

nen diferentes características lingüísticas. Sin embargo, una intuición común sería que la complejidad morfológica depende del sistema morfológico de una lengua, esto es, de sus procesos flexivos y derivativos que permiten la creación de diferentes formas de palabras. Un sistema altamente productivo, producirá muchas palabras distintas. Como esta productividad se ve reflejada en el TTR medido sobre un corpus, TTR se ha utilizado para estimar la complejidad morfológica de las lenguas usando corpus relativamente pequeños. La idea es que un TTR alto corresponde a mayor complejidad morfológica. A pesar de ser muy simple, TTR ha mostrado una correlación alta con mediciones de complejidad más profundas que toman en cuenta características tipológicas extraídas de grandes atlas lingüísticos (Kettunen, 2014; Bentz et al., 2016).

Existen diversos modelos que se han desarrollado para examinar la relación que existe entre tipos y tokens en un texto (Mitchell, 2015). El más común es el cociente $\frac{tipos}{tokens}$, que es el que utilizaremos en este trabajo.

Es importante notar que el valor de TTR se ve afectado por la longitud y tipo de textos. Existen varias alternativas para lidiar con este problema, por ejemplo, normalizar el tamaño del texto o utilizar escala logarítmica. Sin embargo, una manera natural de hacer comparables los TTR de diferentes lenguas es utilizar un corpus paralelo. Por ejemplo, Kelih (2010) trabaja con textos paralelos de lenguas eslavas y analiza características tipológicas de las lenguas, esto es, utiliza TTR para contrastar la productividad morfológica y el grado de síntesis y analiticidad entre las lenguas. En la misma linea, Mayer et al. (2014) extrae automáticamente características tipológicas como el grado de síntesis de la morfología de las lenguas, usando TTR.

Decidimos utilizar TTR para contrastar la complejidad morfológica español-náhuatl y analizar cómo cambia esta complejidad cuando diferentes procedimientos de normalización morfológica son aplicados a los textos (lematización, stemming, segmentación morfológica). Esta información puede ser útil para medir el impacto que tienen diferentes herramientas morfológicas al normalizar la variación de las lenguas. Por otro lado, nuestra intuición es que minimizar el TTR facilitaría la extracción de correspondencias léxicas bilingües y otras tareas de PLN, especialmente en un entorno de bajos recursos.

Para explorar la variación de complejidad entre diferentes tipos de representaciones morfológicas de las lenguas, utilizamos diferentes herramientas. En el caso del español, realizamos lematización, stemming y segmentación morfológica. Para el náhuatl, únicamente segmentación morfológica.

En PLN, el tratamiento de la morfología consiste usualmente en construir analizadores morfológicos, etiquetadores y, más comúnmente, métodos de lematización y stemming para reducir la variación morfológica en un texto llevando todas las formas de palabra a su forma estándar, esto es, un lema o un stem. Sin embargo, la mayoría de estas tecnologías se enfocan en un conjunto reducido de lenguas. En lenguas como el inglés, con una gran cantidad de tecnologías del lenguaje y una morfología relativamente pobre, el tratamiento morfológico automático puede considerarse un problema resuelto.

Sin embargo, este no es el caso para todas las lenguas. Especialmente para las lenguas con morfología rica para las cuales no es suficiente remover las terminaciones flexivas para obtener una raíz o stem.

La lematización y el stemming buscan remover las terminaciones flexivas de las palabras, para el español hay varias tecnologías para hacer esto de manera automática. En este trabajo utilizamos el stemming de Porter (Porter, 2001) así como la herramienta Freeling[9]. La segmentación morfológica del español no es una tarea común, sin embargo, existen algunos métodos tanto supervisados como no supervisados que se han probado exitosamente en esta lengua (Medina-Urrea, 2000; Monson et al., 2004; Gelbukh et al., 2008; Méndez-Cruz et al., 2016). Decidimos utilizar nuestro modelo de segmentación generado con Morfessor, presentado en la sección anterior, pues obtiene resultados equiparables o mejores que estos trabajos previos.

En cuanto al náhuatl, ya hemos comentado que no es fácil encontrar tecnologías del lenguaje, por ejemplo, lematizadores, stemmers, etiquetadores de categorías gramaticales (POS), etc. Utilizamos el modelo de segmentación morfológica semisupervisada, presen-

[9]http://nlp.lsi.upc.edu/freeling/

tado en la sección anterior.

Calculamos la relación tipo token para las dos lenguas utilizando el corpus paralelo. La tabla 4.8 muestra el TTR de los textos sin ningún tipo de procesamiento (ES, NA) y con diferentes tipos de procesamiento morfológico: segmentación morfológica (ES_{morph}, NA_{morph}), lematización (ES_{lemma}) y stemming (ES_{stem}). Los valores de TTR están expresados en porcentaje.

Además, calculamos la diferencia de los TTR obtenidos entre las dos lenguas. La tabla 4.9 muestra la diferencia de TTRs entre los textos de español y náhuatl usando diferentes tipos de normalización morfológica.

	Tokens	**Tipos**	**TTR (%)**
ES	118364	13233	11.17
NA	81850	21207	**25.90**
ES_{morph}	189888	4369	2.30
NA_{morph}	175744	2191	**1.24**
ES_{lemma}	118364	7599	6.42
ES_{stem}	118364	8244	6.96

Tabla 4.8: TTR en diferentes tipos de procesamiento morfológico

Se puede observar que cuando no se aplica ningún procesamiento morfológico, el náhuatl tiene un TTR mucho más alto que el español, esto es, una mayor proporción de diferentes formas de palabra (tipos). Los textos en náhuatl tienen menos tokens que su equivalente en español, esto es esperable pues la naturaleza aglutinante del náhuatl permite que la morfología codifique muchas funciones que se manifiestan en la sintaxis en el español. A pesar de que el náhuatl tiene menos tokens, tiene muchos más tipos que el español. Esto sugiere que el náhuatl posee un sistema altamente productivo, capaz de generar una gran cantidad de diferentes formas morfológicas. En otras palabras, es más probable encontrar una forma de palabra repetida en el corpus de español que en el de náhuatl.

	Diff TTR (%)
$ES_{morph} - NA_{morph}$	1.05
$ES_{lemma} - NA_{morph}$	5.17
$ES_{stem} - NA_{morph}$	5.71
$ES - NA_{morph}$	9.93
$ES - NA$	14.72
$ES_{stem} - NA$	18.94
$ES_{lemma} - NA$	19.48
$ES_{morph} - NA$	23.60

Tabla 4.9: Diferencia de TTRs entre representaciones de las dos lenguas

En todos los casos cuando algún tipo de procesamiento morfológico fue aplicado, el TTR decreció, lo cual puede ser interpretado en términos de reducción de complejidad morfológica de la representación de los textos. Para el español, los textos lematizados obtienen un valor más bajo de TTR que el obtenido con stemming. Aparentemente se alcanza una mejor normalización morfológica usando lemas (se obtienen menos tipos usando lematización que stemming).

La segmentación morfológica indujo los valores más pequeños de TTR para las dos lenguas, sugiriendo que la mayor reducción de complejidad morfológica se alcanza cuando las palabras son segmentadas en morfos, esto es intuitivo: al tener textos segmentados morfológicamente es más probable encontrar una forma repetida. De hecho, cuando el náhuatl es segmentado morfológicamente, el TTR sufre un dramático decremento (del 25.90 a 1,24). Esta gran reducción de TTR puede ser el resultado de haber eliminado la variedad combinatoria de la morfología con tendencia aglutinante y polisintética del náhuatl. Al segmentar los textos, rompemos esta aglutinación provocando unidades menos diversas.

Complementando lo anterior, el hecho de que los textos segmentados hayan reducido su número de tipos e incrementado su número de tokens, enfatiza los beneficios del trata-

miento morfológico para expandir el vocabulario, especialmente en casos de bajos recursos como el náhuatl.

Respecto a las diferencias de TTR entre lenguas (tabla 4.9), la menor diferencia se obtiene cuando las dos lenguas son segmentadas morfológicamente. Esto es interesante, pues el español y el náhuatl tenían originalmente complejidades morfológicas distantes, sin embargo, cuando se segmenta a las dos lenguas o se aplica alguna normalización morfológica al español, las representaciones de las lenguas parecen ser más comparables, lo que puede tener un impacto positivo en la extracción de léxico bilingüe. Si contrastamos las lenguas segmentadas, el náhuatl tiene considerablemente menos morfos que el español. Conjeturamos que esto puede estar relacionado con el hecho de que el español tiene mayor supletividad, o formas "irregulares", en su sistema morfológico. De hecho, existen estudios lingüísticos que sugieren que la supletividad es común en lenguas flexivas, mientras que es rara en lenguas aglutinantes (Veselinova, 2006; Dressler, 1985).

Por otro lado, la máxima diferencia de TTR se obtiene cuando los textos en español son segmentados morfológicamente y a los de náhuatl ningún procesamiento morfológico es aplicado. Esto tiene sentido, pues en este escenario estamos comparando al náhuatl, que por su naturaleza aglutinante tiene menos tokens, contra una representación segmentada del español que provoca aún más tokens de los que ya tenían los textos originalmente.

A este tipo de medidas de complejidad morfológica se les clasifica como basadas en corpus, pues recaen en analizar la distribución de tipos sobre tokens en la producción real de la lengua plasmada en un corpus. Existen otras medidas de este tipo basadas en entropía, sin embargo, es importante tener en cuenta que, a nivel lingüístico, las dimensiones que intervienen en la complejidad de una lengua son muchas, y este tipo de medidas, como la presentada, solo alcanzan a capturar uno de los muchos matices de la complejidad.

4.3. Construcción de lexicón semilla bilingüe

Una vez que realizamos la segmentación morfológica y exploramos la representación del texto que resulta más conveniente, podemos enfocarnos en la tarea de extracción léxica

bilingüe. Como ya hemos visto, lo anterior se realiza usualmente a partir de corpus paralelos, pero también se pueden utilizar corpus comparables o incluso a partir de grandes cantidades de corpus monolingües no relacionados para cada una de las lenguas (inducción de lexicón bilingüe). En nuestro caso tenemos un corpus paralelo español-náhuatl. En PLN, si se tiene un corpus paralelo un paso natural es realizar un sistema de traducción automática. Sin embargo, como en esta tesis nos enfrentamos a un escenario de bajos recursos para lenguas distantes, decidimos tomar un enfoque "first-things-first"(Monson et al., 2006), que establece que cuando se quiere construir sistemas de PLN para lenguas de bajos recursos, es mejor enfocarse primero en producir recursos básicos y, a partir de ello, recursos más complejos. Por lo tanto, en vez de enfocarnos en la traducción automática, en esta tesis se explora qué tan difícil es encontrar correspondencias a nivel palabra así como el impacto de diferentes representaciones morfológicas del texto para mejorar esta tarea.

Uno de nuestros intereses iniciales fue hacer uso de características distribucionales, y de manera más general, el uso de vectores que codifiquen diversas características de las lenguas, para encontrar pares de traducción entre el español y el náhuatl. Ya en la sección 3.1 explicamos como los métodos que explotan características contextuales requieren de algún lexicón semilla o diccionario bilingüe, relativamente pequeño, para poder comparar las representaciones vectoriales contextuales de cada lengua. La mayor parte de los métodos hacen uso de diccionarios bilingües digitales disponibles para un par específico de lenguas. Nuestra intención fue tratar de prescindir de este tipo de diccionarios pre-compilados y, en vez de esto, tratar de inducir un lexicón semilla de manera no supervisada.

Una de las principales motivaciones para realizar lo anterior fue que, aunque existen diccionarios español-náhuatl, no todos ellos están disponibles en versión digital y procesables por la computadora, los más extensos fueron hechos hace siglos y puede haber casos en que las entradas tanto en español como en náhuatl no correspondan con los dialectos hablados hoy en día. Además de la gran variación dialectal y ortográfica que dificulta tener un solo diccionario estándar.

En este sentido, nuestra propuesta de extracción léxica bilingüe se puede plantear como un procedimiento donde intervienen dos etapas. Por un lado queremos hacer una preselección de candidatos a traducción para formar un lexicón semilla, y por otro, construir representaciones vectoriales por palabra y aprender automáticamente un mapeo entre lenguas a partir del lexicón semilla.

Para implementar la primera etapa, exploramos distintos enfoques y características que nos permitieran construir de manera automática un primer lexicón semilla. En particular, nos enfocamos en la combinación de dos métodos estadísticos que son capaces de extraer candidatos a traducción a partir de un corpus paralelo. En los siguientes subcapítulos abordaremos las características de estos métodos y cómo combinarlos para inducir un lexicón semilla confiable.

4.3.1. Método estimativo: IBM-1

En el capítulo 2 se mencionó que los enfoques para realizar alineación de palabras o extracción léxica bilingüe, se pueden clasificar como de tipo estimativo o de tipo asociativo. El enfoque de tipo estimativo surgió a raíz de los modelos estadísticos de traducción automática, se les conoce como estimativos porque los parámetros del modelo de traducción léxica se estiman a través de un proceso de maximización. Esto es, las probabilidades de traducción entre palabras se aprenden a partir de un corpus paralelo alineado a nivel oración, mediante un algoritmo iterativo conocido como esperanza-maximización. El algoritmo se inicializa con una distribución probabilística uniforme, de manera que todas las alineaciones entre palabras son igualmente probables, estas probabilidades se van refinando observando en el corpus paralelo qué pares de traducción co-ocurren en más oraciones paralelas. Lo anterior se logra utilizando estimación de máxima verosimilitud (maximum likelihood estimation) e iterando entre dos pasos hasta converger, el resultado son tablas de traducción léxica que contienen la probabilidad de que una palabra en la lengua destino se traduzca como cierta palabra en la lengua origen $t(e|f)$.

Este planteamiento fue establecido por primera vez en el modelo IBM-1 y después fue escalando hasta llegar al modelo IBM-6 (Brown et al., 1993; Koehn, 2009). Estos modelos subsecuentes agregaron restricciones para refinar los alineamientos entre palabras,

por ejemplo, tomar en cuenta la posición de las palabras para estimar la probabilidad de traducción, fertilidad (una palabra en una lengua puede alinearse con muchas de la otra lengua), reordenamiento basado en combinaciones específicas de clases de palabras, entre otras características. De tal manera que los modelos aumentaron en sofisticación pero también en complejidad y costo computacional.

Elegimos utilizar el modelo IBM-1 pues ha probado ser, aún hoy en día, una referencia o *baseline* fuerte en la tarea de encontrar correspondencias léxicas bilingües (Levy et al., 2016). Además de su bajo costo en términos de complejidad computacional, el hecho de que no tenga restricciones de tipo sintáctico, como los modelos de mayor nivel, puede beneficiar a pares de lenguas como el español-náhuatl, donde no siempre comparten el orden sintáctico y el náhuatl tiene un orden más libre que el español.

La tabla 4.10 muestra un ejemplo de candidatos a traducción obtenidos con el método IBM-1 aplicado al corpus español-náhuatl. La cantidad de candidatos por palabra es variable; en este ejemplo solo seleccionamos los primeros candidatos, es decir, las palabras que obtuvieron mayor probabilidad de ser traducción. t(e|f) indica las probabilidades de traducción léxica entre la lengua origen (e) y la lengua destino (f) como se explicó en el capítulo 2. Es interesante notar que los dos candidatos con mayor probabilidad, corresponden a traducciones correctas: *xochitl* es un sustantivo en náhuatl cuya traducción más común es flor. En realidad, la raíz nominal es *xochi*, *-tl* es un morfo gramatical, marca de caso absolutivo. Vemos que en la tabla aparece también como candidato la forma *xochi*. La presencia de las dos formas se debe a la manera en que nuestro modelo de Morfessor segmentó el texto. Podemos intuir que segmentó correctamente la raíz xochi en ciertos contextos de morfos, mientras que en otros, infirió que -tl era parte de la raíz.

4.3.2. Método basado en asociación: Anymalign (basado en submuestreo)

Las aproximaciones basadas en asociación (a veces también llamados métodos heurísticos) utilizan medidas de similitud o asociación del algún tipo, de manera que se pueda calcular un score que determine qué pares de palabras son candidatos a traducción. Existe una gran variedad de métodos que utilizan este tipo de aproximación, como se discutió

español	**náhuatl**	**t(e|f)**
flor	xochitl	0.5
flor	xochi	0.4
flor	nepapan	0.1
flor	molini	0.09
flor	quitta	0.08
flor	nexti	0.07
flor	itz	0.05
flor	caxtillan	0.04
flor	quitqui	0.03
flor	yan	0.03

Tabla 4.10: Ejemplo de tabla de traducción léxica obtenida con modelo IBM-1

en el capítulo 3.1).

En particular, decidimos utilizar un método de tipo asociativo basado en submuestreo o *sampling-based*, que ha probado tener resultados satisfactorios(Lardilleux y Lepage, 2009; Lardilleux et al., 2011). La idea general de este método es que solo aquellos pares de palabras que ocurren en el mismo par de enunciados paralelos, son considerados para ser candidatos a traducción entre las dos lenguas. Sin embargo, para robustecer estadísticamente al método, se producen artificialmente más candidatos a partir de formar muchos subcorpus de tamaño pequeño (sub-muestreo) extraídos del corpus paralelo original.

Este método recibe como entrada un corpus paralelo alineado a nivel oración y devuelve pares de palabras (candidatos de traducción) con un score global que refleja qué tan fuerte es la asociación entre el par de palabras. El planteamiento del método de solo tomar en cuenta pares de palabras que ocurren en las mismas oraciones paralelas, podría parecer contrario al utilizado por la mayor parte de los métodos que prefieren palabras de frecuencia alta y que, incluso, muchas veces descartan las que no superan un umbral de frecuencia. Los autores arguyen que tratar de alinear palabras de menor frecuencia debería ser más sencillo que alinear las de alta frecuencia. Por ejemplo, las palabras que solamente aparecen una vez, happax legomena o singletons, se vuelven el caso más fácil

de alinear si, además de ocurrir una sola vez en cada lengua, aparecen en la misma oración paralela. En otras palabras, determinar si dos palabras frecuentes que ocurren en más o menos las mismas oraciones son traducción, es un problema de decisión con varias opciones de solución, mientras que para los singletons no hay necesidad de aproximar la solución.

Por lo tanto, este método se enfoca en pares de palabras que comparten estrictamente la misma distribución en las oraciones del corpus paralelo, sin importar su frecuencia. Sin embargo, como este tipo de escenarios es raro y quizá pocas palabras del corpus paralelo cumplirían esta restricción, los autores proponen una forma de generar más candidatos de traducción y de que ganen significancia: construir pequeños subcorpus paralelos a partir del original y extraer de ahí candidatos a traducción. A este proceso le llaman sub-muestreo, pues se obtienen muchos subcorpus del corpus paralelo original, a través de un muestreo, y se van procesando estos extractos uno tras otro. Entre más pequeño sea el corpus paralelo, menor es la frecuencia de las palabras en las dos lenguas y, por lo tanto, es más probable que compartan la misma distribución en el texto. Esta forma de ir seleccionando aleatoriamente muestras de enunciados paralelos del corpus paralelo, se puede ver como una forma de producir más candidatos "artificialmente". Finalmente, las palabras o secuencias de palabras que comparten la misma distribución son extraídas para formar una tabla de pares de traducción junto con el número de veces que fueron alineadas durante el muestreo.

Para implementar el método basado en submuestreo con el corpus paralelo español-náhuatl, utilizamos una herramienta de software libre llamada Anymalign[10] que permite obtener las correspondencias léxicas a nivel palabra e incluso multi-palabra. El número de subcorpus aleatorios que se generan no se establecen a priori, de manera que el proceso se puede interrumpir en cualquier momento. Entre más tiempo corra el programa, se procesan más subcorpus y se obtienen más candidatos de traducción. La tabla 4.11 contiene un ejemplo de los primeros pares de traducción (mayor score global) obtenidos con este método, a diferencia del ejemplo de la tabla 4.10 con el método IBM, donde ordenamos las probabilidades de manera local para cada palabra.

[10]http://anymalign.limsi.fr

Podemos observar que este método permite obtener correspondencias léxicas a nivel multipalabra, como en el caso de *frase:tlahtol tin*, *dialogar:nonotza liztli*. En realidad *-tin* y *-liztli* corresponden a sufijos nominales (segmentados correctamente), el método los asocia como traducción de una sola palabra pues suelen aparecer juntos.

español	**náhuatl**	**score**
el	in	164651
vender	tlanamaca	97743
y	ihuan	84932
dinero	tomin	79051
no	amo	78881
vivir	nemi	68006
frase	tlahtol tin	65469
aquí	nican	58458
amar	tlazohtla	58446
este	inin	49838
más	ocachi	49349
dialogar	nonotza liztli	46828
obispo	obispo	44653
dios	dios	42196
yo	nehua	41559

Tabla 4.11: Ejemplo de tabla de traducción léxica obtenida con el método basado en submuestreo (anymalign)

4.3.3. Combinación de métodos

En áreas como la recuperación de la información y el aprendizaje automático supervisado, la combinación de métodos o tomadores de decisiones que trabajen juntos suelen encontrar mejores soluciones que trabajando solos. Dietterich (1998) establece que una

combinación exitosa de métodos requeriría que los métodos tengan un desempeño comparable y que además no cometan errores similares. Debido a esto, elegimos combinar los dos métodos presentados en las secciones anteriores para formar un lexicón semilla, ya que tienen desempeños comparables, pero al estar basados en aproximaciones distintas probablemente no cometan el mismo tipo de errores.

Es importante mencionar que no es común probar estos métodos en entornos de bajos recursos, sin embargo existen algunos antecedentes, por ejemplo, el método basado en submuestreo (Anymalign) ha mostrado buena exactitud en la extracción de candidatos de traducción para palabras de baja frecuencia (Kwon et al., 2014; Lardilleux et al., 2011). Por otro lado, los modelos IBM han sido utilizados para extraer lexicón bilingüe a partir de pequeñas cantidades de corpus paralelo formado por transcripciones fonémicas (Adams et al., 2015).

Creamos nuestro lexicon semilla, combinando el método estimativo (IBM-1) y el asociativo (Anymalign) de la siguiente forma:

1. Para cada método, se extraen todos los pares de traducción que son simétricos, esto es, los pares de traducción que coinciden cuando se aplica el método tanto de español a náhuatl como de náhuatl a español. Esta restricción de simetría puede ayudar a obtener pares de traducción altamente confiables (Vulic y Korhonen, 2016).

2. Se conservan estos pares simétricos para cada uno de los métodos y se elabora una lista rankeada. En primer lugar se coloca el subconjunto de pares simétricos que coinciden en los dos métodos, después se enlistan el resto de pares simétricos obtenidos con cada uno de los métodos

3. Finalmente se realiza un corte arbitrario para filtrar los primeros n pares de la lista rankeada. Este conjunto de pares de palabras constituye nuestro lexicón semilla

Nuestra intuición general fue que aquellos candidatos a traducción que son simétricos y que coinciden utilizando más de un método, pueden ser altamente confiables para formar un lexicón semilla sin necesidad de intervención de un humano. Este lexicón se utilizará para aprender una transformación lineal entre los espacios semánticos de las dos lenguas como se explicará en el siguiente subcapítulo.

4.4. Representaciones vectoriales para extracción léxica bilingüe

Uno de nuestros objetivos principales es averiguar qué características o información se pueden combinar para realizar una estimación adecuada de pares de traducción español-náhuatl a partir del corpus paralelo. Hasta ahora hemos combinado aproximaciones de tipo estimativo y asociativo para hacer una pre-selección de candidatos a traducción. Uno de los aspectos en los que tenemos especial interés es en utilizar representaciones vectoriales y explotar información contextual: una palabra que ocurre en cierto contexto en una lengua, debe tener una traducción que ocurra en un contexto similar en la otra lengua.

En el capítulo 2.4 se explicó que existen representaciones vectoriales que codifican el significado de una palabra por medio de sus contextos de aparición. Los métodos de extracción léxica bilingüe que utilizan información de tipo contextual, necesitan comparar los vectores pertenecientes a cada una de las lenguas para medir cuáles están más cercanos y, por lo tanto, representan un par de traducción. Para realizar lo anterior existen diversas aproximaciones que buscan, en su mayoría, hacer comparables los vectores por medio de una señal supervisada (lexicón semilla) que permite aprender una transformación o proyección. En particular, en este trabajo estamos interesados en aprender una transformación lineal que relacione los espacios vectoriales del español y el náhuatl, como se explicará en la sección 4.4.2.

En cuanto a las representaciones vectoriales, hoy en día en PLN se utilizan ampliamente las de tipo distribuido o word embeddings para una gran gama de tareas, incluidas las tareas multilingües. En este sentido, utilizaremos las representaciones distribuidas Word2Vec (w2v), basadas en modelos neuronales del lenguaje como se explicó en la sección 2.4.1. Utilizamos nuestro corpus paralelo para entrenar representaciones W2V monolingües.

Sin embargo, una de las limitaciones de este tipo de representaciones contextuales es la cantidad de corpus requerido para obtener representaciones vectoriales de buena calidad. Entre más repeticiones de una palabra en el corpus, mayor el número de contextos

en los que aparece la palabra y mejor la calidad del vector de contextos que captura su significado. Por lo tanto, cuando este tipo de vectores es generado a partir de corpus pequeños, como el nuestro, las representaciones pueden no capturar propiedades semánticas y no tener buen desempeño al utilizarlo en alguna tarea de PLN.

Debido a lo anterior, decidimos proponer una forma de generar representaciones vectoriales multilingües para el español y el náhuatl que pudieran ser útiles para la tarea de encontrar pares de traducción en un entorno de bajos recursos. Estas representaciones se abordan en la siguiente sección 4.4.1

4.4.1. Representaciones distribuidas multilingües basadas en grafos

Como ya hemos mencionado, las representaciones vectoriales distribuidas, como W2V, son representaciones densas (no dispersas) de valores reales que codifican el significado de una palabra o alguna unidad lingüística por medio de sus contextos. Desafortunadamente, enfrentamos un entorno experimental de pocos recursos disponibles. Los métodos de representaciones distribuidas o word embeddings, necesitan una cantidad grande de datos para alcanzar una distribución probabilística adecuada en el espacio vectorial (Bengio et al., 2003; Mikolov et al., 2013a).

Con el objetivo de lidiar con este problema y obtener representaciones que puedan ser de utilidad para nuestra tarea, proponemos una alternativa para construir representaciones vectoriales de tipo multilingüe. Nuestra idea parte del hecho de que tenemos un corpus paralelo que, aunque pequeño, tiene anotadas sus correspondencias a nivel oración o párrafo. Esta información es suficiente para obtener candidatos de traducción a nivel palabra con métodos estadísticos como los utilizados en las secciones previas. Este tipo de métodos arroja para cada palabra una lista de candidatos a traducción asociados a un score o probabilidad. Estos resultados pueden interpretarse como una estructura de grafo, nuestra idea es aprovechar esta estructura de grafos y a partir de ella generar representaciones distribuidas bilingües.

Existen algoritmos para aprender representaciones vectoriales continuas a partir de un

grafo, en particular utilizamos Node2vec (Grover y Leskovec, 2016). Esta formulación es similar a Word2Vec, sin embargo, este método toma como entrada un grafo $G = (V,E,\phi)$ donde V es el conjunto de nodos, E es el conjunto de arcos y $\phi : E \to \mathbb{R}$ es una función que establece el peso de cada arco. El objetivo de Node2Vec es encontrar un mapeo $F : V \to \mathbb{R}^d$, transformando así los nodos del grafo en vectores de una dimensión d, al igual que en Word2Vec este parámetro se puede variar.

De manera muy general, este problema se plantea como un problema de aprendizaje de minimización de riesgo por máxima verosimilitud (maximum likelihood risk minimization) que puede definirse de la siguiente manera (Vapnik, 1998):

$$Q(x,\theta) = -\log q(x|\theta) \tag{4.1}$$

Donde x representa los datos y θ los parámetros de aprendizaje. Node2Vec toma una vecindad $N_S(v) \subseteq V$ de un nodo $v \in V$ utilizando una estrategia de muestreo (que se basa en caminatas aleatorias en el grafo). Si $f(v) \in \mathbb{R}^d$ es la representación vectorial de un nodo $v \in V$, el problema de máxima verosimilitud en Eq. 4.1 puede ser replanteado con esta función objetivo a optimizar:

$$\max_f \sum_{v \in V} \log p(N_S(v)|f(v)) \tag{4.2}$$

Finalmente el método devuelve un vector por cada nodo del grafo, este vector codifica la vecindad del nodo. Es importante destacar que una de las bondades del algoritmo de Node2Vec recae en el hecho de que puede realizar caminatas aleatorias de segundo orden. Una caminata aleatoria de segundo orden es cuando la decisión de cuál es el siguiente nodo a seguir no se basa solamente en el nodo actual, sino, también en el nodo anterior. Este tipo de caminatas permiten una mayor exploración del grafo. Finalmente, estas relaciones más profundas entre nodos pueden impactar en el tipo de representaciones vectoriales obtenidas. Node2vec permite controlar este y otros parámetros de las caminatas aleatorias.

En este trabajo, construimos un grafo a partir de los candidatos de traducción y su score obtenido con el método basado en submuestreo Anymalign. De manera más formal, sean $w^1 \in L_1$ y $w^2 \in L_2$ el conjunto de palabras pertenecientes a la lengua origen y a la lengua destino, entonces $(w^1,w^2) \in E$ si w^2 es una posible traducción de w^1 de acuerdo

al método Anymalign. Por lo tanto, la función de pesos del grafo $\phi(w^1, w^2)$ es obtenida a partir de un método asociativo de alineación de palabras aplicado al corpus paralelo español-náhuatl. Es importante mencionar que elegimos el método anymalign por su bajo costo computacional y buenos resultados (capítulo 5), sin embargo, la distribución del grafo podría ser inicializada con cualquier otra metodología que relacione la palabras entre las dos lenguas.

Una vez construido el grafo $G = (V,E,\phi)$, fue posible generar una representación vectorial para cada nodo aplicando Node2Vec. Los detalles de los parámetros utilizados se explicarán en el capítulo 5 de evaluación.

La figura 4.3 muestra vectores de Word2Vec para algunas palabras del corpus paralelo. Los triángulos apuntando hacia abajo (azules) corresponden a español, los triángulos apuntando hacia arriba (rojos) a náhuatl. En paréntesis se pone la traducción a inglés solo con fines ilustrativos

La figura 4.4 muestra las mismas palabras pero ahora usando nuestra representación propuesta usando grafos y Node2Vec. Igualmente los triángulos apuntando hacia abajo (azules) corresponden a español, los triángulos apuntando hacia arriba (rojos) a náhuatl. En todos los casos, para poder realizar las gráficas de los vectores en dos dimensiones se utilizó la técnica t-SNE (Hinton y Roweis, 2003; Maaten y Hinton, 2008).

Aunque se trata de un ejemplo simbólico que involucra solo unas cuantas palabras. En las representaciones gráficas podemos observar que los vectores Word2Vec entrenados de manera monolingüe tienen cierta distribución en el espacio: los vectores pertenecientes a náhuatl se separan de los vectores pertenecientes al español. Esto es esperable, sin embargo no se ve claro si existiría una transformación lineal que pueda relacionar a cada palabra con su traducción correspondiente.

Por otro lado, nuestras representaciones vectoriales bilingües, obtenidas con Node2Vec, muestran un comportamiento diferente. Se puede observar que nuestro planteamiento provocó que los vectores de cada palabra de la lengua fuente estén cercanos al vector

de su traducción correspondiente en la lengua destino. Asimismo, la forma en cómo se distribuyen estos vectores en el espacio, también permiten conjeturar que utilizando este tipo de representaciones será más fácil encontrar una transformación lineal que refine, o acerque más, a los pares de traducción.

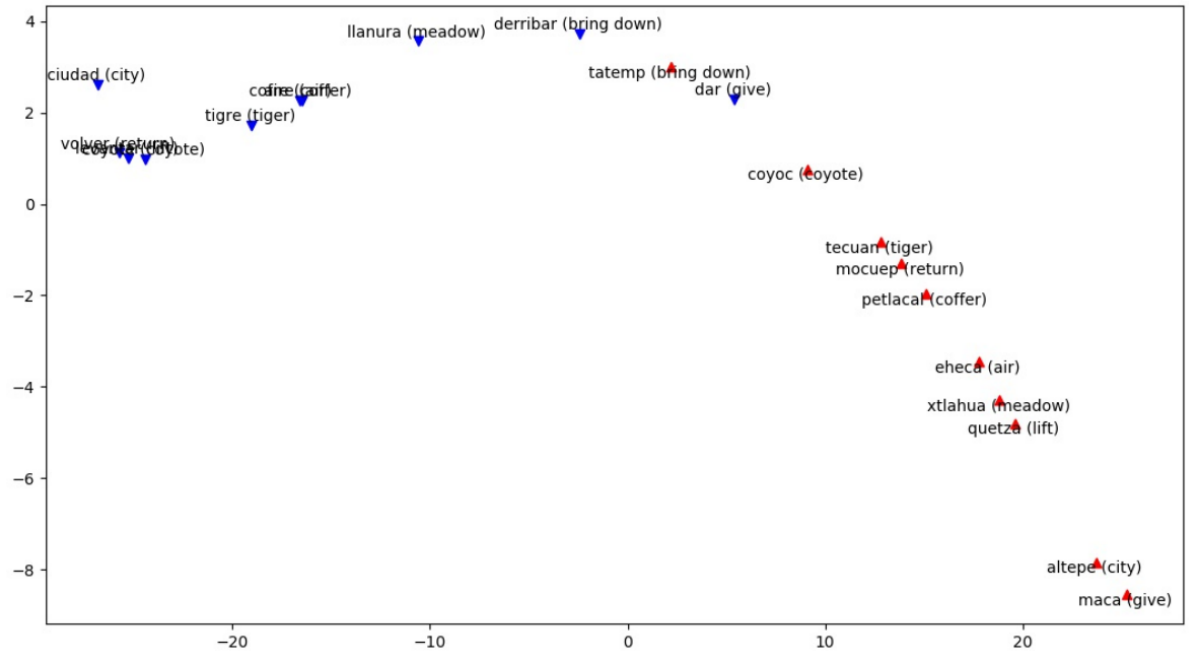

Figura 4.3: Vectores de palabra usando Word2Vec

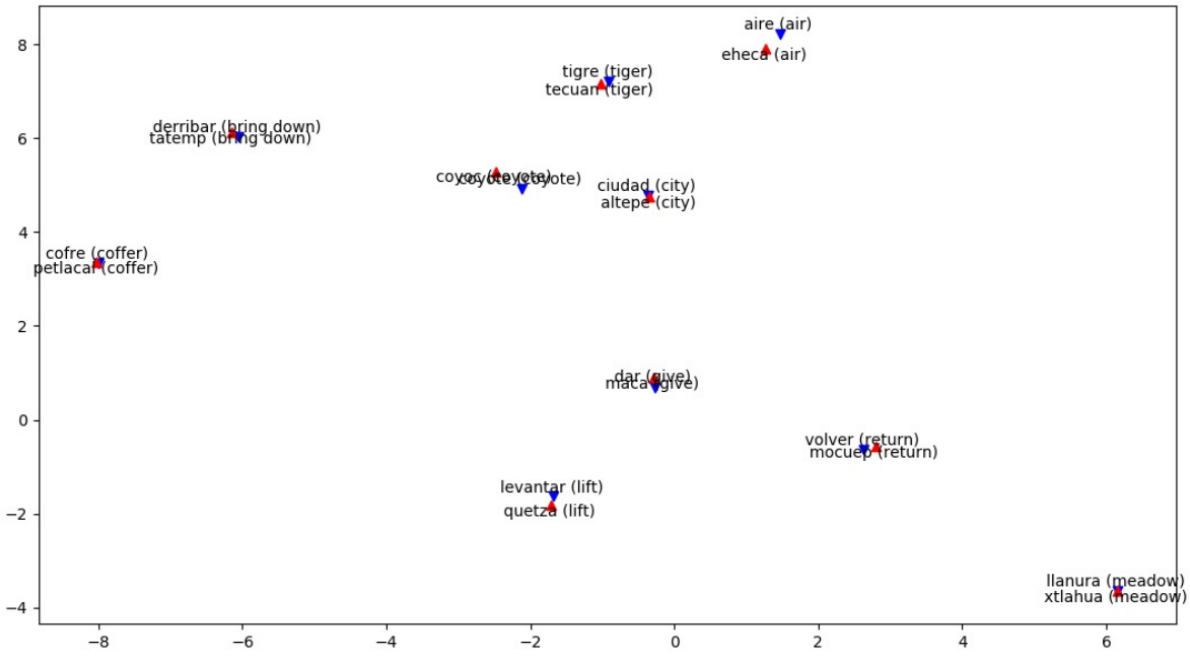

Figura 4.4: Vectores de palabra usando Node2Vec

4.4.2. Mapeo lineal entre lenguas

Ya hemos comentado que los métodos de extracción léxica bilingüe que utilizan representaciones vectoriales a nivel palabra, tratan de construir representaciones multilingües o de mapear los vectores monolingües de cada lengua a un espacio común en donde se puedan encontrar las traducciones más cercanas (Lauly et al., 2014; Hermann y Blunsom, 2014; Mikolov et al., 2013b). Estos métodos del estado del arte no necesitan forzosamente de corpus paralelo, son capaces de inducir léxico bilingüe a partir de cantidades enormes de corpus monolingüe o comparable. Sin embargo, también se ha demostrado que cuando enfrentan un escenario de bajos recursos, pueden tener incluso peor desempeño que métodos menos sofisticados (Levy et al., 2016).

Nuestro objetivo es enfocarnos en el tipo de métodos que encuentran un mapeo lineal que relacione los espacios de dos lenguas, sin embargo, proponemos adaptarlo para que funcione bien en entorno de bajos recursos, esto es, utilizando las representaciones bilingües vectoriales basadas en grafos que propusimos en la sección anterior.

Mikolov et al. (2013b) propuso una manera de encontrar una transformación lineal que mapea vectores de una lengua hacia otra. En el experimento original, los espacios vectoriales de cada lengua se inducen con grandes corpus monolingües y utilizando representaciones distribuidas Word2Vec generadas con la arquitectura CBOW o con Skip-gram.

Para poder aprender este mapeo lineal entre los dos espacios se necesita en primer lugar una señal supervisada. Esto es, un lexicón semilla que contenga una lista de pares de traducción correctos entre las dos lenguas. Los pares de traducción se pueden expresar como (x_i, y_i), donde $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}^{d_2}$

La idea es que, a partir de este lexicón semilla, podemos explotar propiedades geométricas de los espacios vectoriales de cada lengua. Mikolov et al. (2013b) notó de manera gráfica que las representaciones vectoriales de las traducciones entre dos lenguas mantienen un arreglo geométrico similar. De hecho, estas regularidades son lineales, por ejemplo, si aplicamos una rotación, una traslación, un escalamiento a los vectores de la lengua

fuente, podemos acercarlos a su traducción correspondiente en la lengua destino.

Por lo tanto, el objetivo es encontrar una transformación lineal $T : \mathbb{R}^d \rightarrow \mathbb{R}^{d_2}$ que tome un vector de la lengua fuente y lo transforme a su traducción en la lengua destino. Si $x_i \in \mathbb{R}^d, i = 1,2,...,N$ son vectores palabra de la lengua fuente y $y_i \in \mathbb{R}^{d_2}, i = 1,2,...,N$ son los vectores correspondientes a posibles traducciones en la lengua destino, entonces T puede aprenderse mediante el siguiente problema de optimización:

$$\min_{T} \sum_{i=1}^{N} ||T(x_i) - y_i||^2 + \lambda||T||^2 \tag{4.3}$$

Este es un problema de mínimos cuadrados que se puede resolver con un método de gradiente descendiente estocástico. La idea es encontrar aquella matriz T (la transformación lineal) que multiplique a los vectores de la lengua fuente, tal que el resultado sean vectores transformados que tengan una distancia mínima a su respectiva traducción. Esta optimización tiene término de regularización $\lambda \in \mathbb{R}$.

El lexicón semilla español-náhuatl generado a partir de la combinación de diferentes métodos (sección 4.3) fue utilizado para aprender la transformación lineal entre los vectores de cada lengua. En resumen, el procedimiento general necesario para realizar nuestro mapeo lineal entre lenguas, involucra estos pasos previos que se abordaron a lo largo del capítulo:

1. Generar un lexicón semilla combinando un método estimativo y uno asociativo de alineación a nivel palabra.

2. Generar representaciones vectoriales multilingües de las palabras en español y náhuatl. Estos vectores se generan a partir de un grafo que es inicializado con los scores de traducción entre pares de palabras (scores obtenidos con un método asociativo basado en submuestreo). Posteriormente, el algoritmo Node2Vec permite convertir cada nodo (palabra) en una representación vectorial continua.

3. Una vez generados los vectores, se aprende una transformación lineal para mapear los vectores del Español hacia el espacio del Nahuatl. Esta transformación es aprendida a partir del lexicón semilla y planteado como un problema de optimización de mínimos cuadrados.

4. La matriz de transformación es aplicada a un conjunto de evaluación de palabras en español. Para cada vector proyectado (español), los vectores vecinos más cercanos (náhuatl) corresponden a los candidatos de traducción.

La tabla 4.12 muestra un ejemplo de un par de traducción, extraído de nuestro corpus, que resultó simétrico con los dos métodos que utilizamos. Al ser un par simétrico que coincide en los dos métodos se pone en los primeros lugares del ranking que formará al lexicón semilla.

IBM	(prob)	ANYM	(score)
cebolla-xonaca	0.65	cebolla-xonaca	173,876
xonaca-cebolla	0.59	xonaca-cebolla	165,144

Tabla 4.12: Ejemplo de par simétrico para formar el lexicón semilla

Capítulo 5

Evaluación y análisis de resultados

5.1. Impacto de la morfología en método estimativo y asociativo

En el subcapítulo 4.2 se discutieron los diferentes procedimientos de normalización morfológica aplicados a los documentos del corpus paralelo y se midió la complejidad morfológica de cada una de las representaciones obtenidas. Lo anterior gira en torno a nuestra conjetura de que lidiar con la morfología puede ser benéfico para mejorar la extracción de correspondencias bilingües en un entorno de bajos recursos, pues lenguas con morfología muy rica son capaces de producir muchas formas de palabra distintas y con menor probabilidad de que se repitan en un corpus, lo que es problemático para los métodos estadísticos. Es por eso que normalizar morfológicamente nos puede llevar a tipos menos diversos con más repeticiones en el texto. Además de permitir alinear unidades léxicas entre dos lenguas distantes con mayor facilidad.

Con el fin de comprobar lo anterior de manera empírica, evaluamos el desempeño de dos métodos de alineación de palabras utilizando los diferentes procesamientos morfológicos del corpus paralelo español-náhuatl. Se usaron dos de los métodos de alineación descritos en la metodología: el método asociativo Anymalign y el método estimativo IBM-1.

Realizamos una evaluación manual seleccionando una muestra aleatoria de 150 palabras en español (no funcionales) con frecuencia mayor a 2. Detallaremos primeramente cómo se construyó el conjunto de evaluación pues este conjunto, con diferentes variaciones, es el que se ocupa para evaluar los diferentes métodos de extracción léxica bilingüe implemen-

tados en esta tesis. A partir de la muestra aleatoria de palabras en español, se construyó un lexicón de evaluación (o prueba) anotando por cada palabra del español una lista de posibles candidatos correctos a traducción en náhuatl.

Construir este lexicón de evaluación no es una tarea trivial, especialmente en un entorno experimental como el nuestro. En trabajos pertenecientes al estado del arte los lexicones de evaluación se extraen de diccionarios bilingües disponibles para el par de lenguas a evaluar. Sin embargo, es importante recordar que en nuestro caso la extracción léxica bilingüe se realizó a partir de un corpus paralelo que presenta variación diacrónica, dialectal y ortográfica del náhuatl. Lo anterior implica que para juzgar si un par de traducción es correcto se necesita que el evaluador tenga conocimiento de las diferentes alternativas de traducción a través de las diferentes variantes del náhuatl y/o en diferentes periodos históricos de la lengua. Es difícil extraer este tipo de información de un sólo diccionario bilingüe español-náhuatl debido a la falta de normalización. Además, nuestro trabajo implica el uso de diferentes representaciones morfológicas (lemas, morfemas, formas conjugadas), característica ausente en otros trabajos, por lo que nuestra evaluación requiere, también, de conocimiento morfológico de las lenguas.

Por las razones anteriores decidimos que un juez humano, con conocimiento lingüístico del náhuatl, fuera quien anotara qué candidatos a traducción podíab considerarse correctos para una palabra. Una evaluación humana es costosa, por el tiempo que implica y la dificultad de encontrar anotadores con el perfil necesario para evaluar la tarea. En nuestro caso esos factores fueron prohibitivos y sólo contamos con un juez humano: la autora de esta tesis. La evaluación estuvo sustentada en la consulta de diferentes materiales lexicográficos del náhuatl, esto es, más de diez diccionarios de diversas épocas (1547 a 2002) todos ellos concentrados en el proyecto del Gran Diccionario Náhuatl (GDN). Asimismo, la evaluación está sustentada en el conocimiento de tipo lingüístico del náhuatl de la evaluadora[1]. Las limitaciones de esta evaluación se discuten más a fondo en la sección de discusión (5.4) y en trabajo futuro (6.2).

[1]Curso de Náhuatl Clásico (análisis lingüístico) concluido en el Instituto de Investigaciones Antropológicas de la UNAM en el periodo 2014-2016. Conclusión de estudios de lengua náhuatl como segunda lengua (periodo 2014-2017) en Centro de Enseñanza de Lengua Extranjera (CELE) UNAM.

Una vez construido este lexicón de evaluación de referencia, se puede evaluar el desempeño de los diferentes métodos automáticos de extracción léxica bilingüe. Por lo tanto, para cada palabra en español se obtuvo una lista rankeada de traducciones en náhuatl utilizando los diferentes métodos de extracción léxica y, también, usando las diferentes representaciones morfológicas propuestas. La idea de la evaluación es verificar si los candidatos a traducción, obtenidos automáticamente, para una palabra coinciden con las alternativas de traducción anotadas manualmente en el conjunto de prueba.

Para asegurar que los resultados fueran comparables a través de las diferentes representaciones morfológicas, tratamos de usar, en la medida de lo posible, el mismo conjunto de palabras de evaluación, esto es, tomamos la muestra aleatoria de palabras en español (sin aplicar ningún tipo de análisis morfológico) y posteriormente las lematizamos y segmentamos morfológicamente para construir los diferentes conjuntos de evaluación. La tabla 5.1 muestra las precisiones $p@1$, $p@5$ obtenidas con cada método aplicado a los textos paralelos con diferentes representaciones morfológicas. Es importante mencionar que la precisión en k (usualmente expresada como $p@k$), es un tipo de precisión comúnmente utilizada en tareas de recuperación de la información donde se tienen que evaluar si el resultado correcto aparece dentro de una lista rankeada. En la tarea de extracción léxica bilingüe, $p@k$ indicaría la proporción de traducciones correctas que fueron encontradas dentro de los primeros k candidatos que arroja el método en cuestión de extracción léxica bilingüe (Schütze et al., 2008).

Se puede observar en la tabla 5.1, que todos los tipos de procesamiento morfológico que aplicamos a los textos, ayudaron a mejorar la extracción léxica bilingüe en nuestro entorno de bajos recursos. La representación más apropiada parece ser ES_{lem}-NA_{morph}, es decir, cuando los textos en español están lematizados y los de náhuatl segmentados morfológicamente. Este tipo de representación no solo provoca la precisión mas alta en la extracción de pares de traducción, sino que los pares obtenidos son más cercanos a una entrada de diccionario, esto es, formas de palabra con pocas o nulas marcas flexivas. A pesar de que la normalización morfológica resulta benéfica para cualquiera de los métodos, el método asociativo basado en muestreo ($ANYM$) parece especialmente beneficiado

del tratamiento morfológico.

Es importante mencionar que la evaluación de ES_{morph}-NA_{morph} no fue tan directa, nos encontramos varios casos problemáticos, por ejemplo, morfos del español que codifican varias funciones gramaticales y para los cuales no era posible encontrar una relación uno a uno; o raíces del español ambiguas, es decir, con la misma forma pero diferente significado (sincretismo morfológico), entre otros fenómenos. En general, tratamos de alinear solo raíces o radicales con significado léxico, por lo que varios pares de traducción devueltos por los métodos tuvieron que ser descartados en la evaluación de esta configuración experimental.

Adicionalmente, estábamos interesados en realizar un mayor análisis a nivel cuantitativo y cualitativo, por lo que nos enfocamos en el análisis de los sustantivos y verbos, en español, contenidos en el conjunto de evaluación. La tabla 5.2 muestra la proporción de verbos y sustantivos que fue correctamente traducida (del total de verbos y sustantivos en los conjuntos de evaluación). En la tabla se muestran las precisiones p@1, p@5 obtenidas con cada método y utilizando diferentes representaciones morfológicas.

	IBM %		ANYM %	
	p@1	p@5	p@1	p@5
ES-NA	48.9	73.1	43.8	61.3
ES_{lem}-NA_{morph}	54.6	78.6	**66.6**	**89.3**
ES_{morph}-NA_{morph}	49	73.9	57.4	79.9

Tabla 5.1: Evaluación de métodos IBM y Anymalign con diferentes representaciones morfológicas

De esta tabla nos gustaría destacar que una vez más las representaciones del texto con algún tipo de procesamiento morfológico provocan una mayor proporción de verbos y sustantivos correctamente traducidos (de un total de 53 verbos y 57 sustantivos). En particular pusimos atención a los verbos, pues representan la clase de palabra morfológicamente más compleja del español y del náhuatl, por lo que esperaríamos que estas

traducciones fueran las más difíciles de obtener correctamente. La representación ES_{lem}-NA_{morph} combinada con el método Anymalign obtiene la mayor proporción de verbos traducidos correctamente.

En un análisis más cualitativo, nos dimos cuenta que algunos verbos del español a

	Verbos %		Sustantivos %	
ES-NA	p@1	p@5	p@1	p@5
IBM	41	66	50.9	75.4
ANYM	33.9	55.3	52.8	67.9
ES_{lem}-NA_{morph}	p@1	p@5	p@1	p@5
IBM	58.3	79.1	50	67.2
ANYM	70.8	**89.5**	53.4	74.1
ES_{morph}-NA_{morph}	p@1	p@5	p@1	p@5
IBM	38.2	61.7	54.3	**78.2**
ANYM	47	79.4	60.8	73.9

Tabla 5.2: Proporción de verbos y sustantivos del español que fueron correctamente traducidos

pesar de ser muy frecuentes, resultaron difíciles de traducir por los métodos automáticos, por ejemplo, verbos copulativos (ser, estar) y auxiliares (haber, ir, tener). Conjeturamos que este tipos de verbos fueron difíciles de alinear porque en náhuatl, al ser una lengua tipológicamente distante, se manifiestan de diferente manera. Por ejemplo, estos verbos que se manifiestan en la sintaxis en el español, en náhuatl suelen ser parte de la estructura morfológica (aparecen aglutinados a sustantivos o verbos); además, en algunos casos pueden no aparecer en absoluto y por lo tanto no tener una correspondencia directa al lema del español.

Finalmente, debido a los buenos resultados obtenidos con la representación morfológica ES_{lem}-NA_{morph}, elegimos utilizarla en todos los experimentos posteriores de nuestra metodología para realizar la extracción léxica español-náhuatl, y de los cuales se hablará en el siguiente subcapítulo.

5.2. Lexicón Semilla

En el capítulo 4 se abordó la metodología propuesta para realizar extracción léxica español-náhuatl. Uno de los primeros pasos lo constituye la inducción de un lexicón semilla que está formado por pares de traducción y que permite aprender una transformación que relacione a los espacios de las dos lenguas.

Como ya se explicó en el capítulo anterior, nuestra propuesta consistió en inducir un lexicón semilla combinando los resultados de dos métodos estadísticos (subcapítulo 4.3). El tamaño del lexicón está determinado por un corte arbitrario en una lista rankeada; de manera específica, nuestro lexicón semilla resultó de 496 entradas español-náhuatl, sin palabras repetidas en español (no más de una traducción por palabra). También, es importante recordar, que las representaciones morfológicas del texto que se utilizan a partir de esta etapa de la metodología son ES_{lem}-NA_{morph}, por lo que el lexicón semilla está formado por lemas del español pareados con morfos del náhuatl.

Si quisiéramos evaluar de manera directa la calidad del lexicón semilla, resultaría costoso evaluar más de 400 entradas manualmente (en la sección de discusión hablaremos sobre las dificultades de la evaluación). Sin embargo, creemos que las restricciones impuestas para su generación aseguran en cierto grado la confiabilidad de los pares de traducción. Por otro lado, de manera indirecta podemos conocer su calidad al analizar qué tanto sirvió este lexicón semilla para aprender una transformación capaz de proyectar vectores de palabras en español a su traducción correcta en náhuatl, como se verá en la siguiente sección en la evaluación de la extracción léxica bilingüe.

A manera de análisis exploratorio, en la figura 5.1 se muestra la distribución de categorías gramaticales (etiquetas POS *part-of-speech*) de las 496 palabras en español. El lexicón semilla resultó estar formado primordialmente por sustantivos y verbos, y en menor medida por otro tipo de clases de palabra. Es importante mencionar que en la tarea de extracción léxica bilingüe final nos centraremos en unidades con significado léxico y filtraremos las de contenido gramatical, pero en el lexicón semilla no las descartamos pues su aparición puede ser de utilidad para ayudar a aprender una transformación bilingüe

más general.

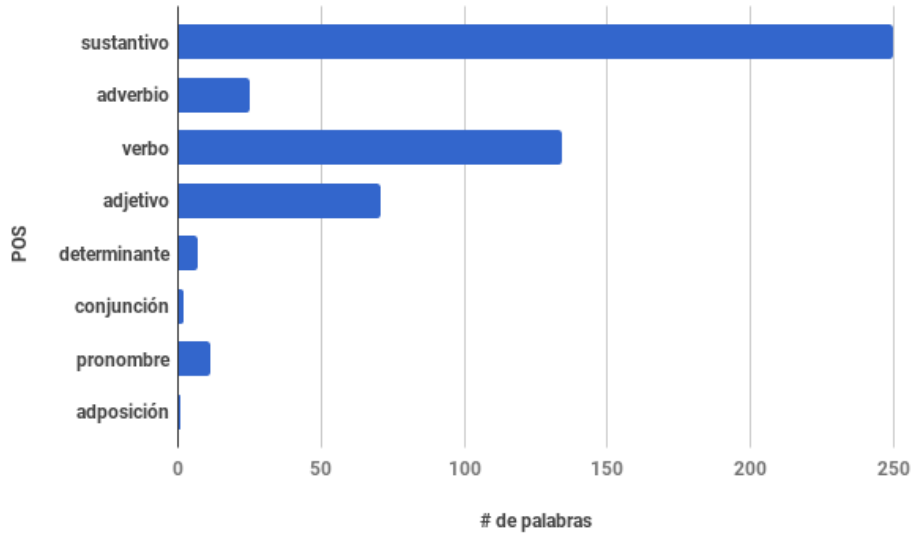

Figura 5.1: Categorías gramaticales del lexicón semilla

5.3. Representaciones vectoriales y mapeo lineal entre español-náhuatl

Una vez inducido el lexicón semilla, el siguiente paso es generar las representaciones vectoriales para cada una de las lenguas y aprender una transformación lineal que las relacione. Como parte de la evaluación contrastamos representaciones vectoriales distribuidas obtenidas con Word2Vec y nuestras representaciones bilingües construidas a partir de anymalign, grafos y Node2Vec. Detallaremos aquí los parámetros utilizados que mejores resultados nos dieron.

Nuestras representaciones vectoriales Node2Vec se obtienen a partir de un grafo que relaciona las palabras de las dos lenguas; esta estructura de datos se obtiene a partir de los resultados del método asociativo basado en submuestreo, Anymalign. Como ya hemos mencionado, este método devuelve una lista de pares de traducción rankeados descendentemente por un score (ejemplo en apéndice E), el número de pares devueltos depende del

tamaño del conjunto de evaluación y de los candidatos de traducción encontrados para cada palabra (que puede ser variable), las especificaciones de los conjuntos de evaluación se abordarán en el siguiente capítulo. Para generar el grafo, almacenamos solo los pares que superaran un umbral de score (score mayor a 1000, aproximadamente resultaron 3000 pares). La motivación de esta decisión responde al costo computacional y también a que entre más bajo sea el score, menor es la confianza en que la traducción sea correcta. De hecho, cuando probamos generar el grafo con la mayor parte de los resultados arrojados por anymalign, nuestros resultados finales de la extracción léxica empeoraron y el costo computacional era significativamente mayor.

El grafo resultante tuvo 5474 nodos, donde cada nodo representa una palabra y los enlaces tienen asociado un peso o score que índica qué tan fuerte es la relación entre dos palabras (traducción), en principio es un grafo bipartito no dirigido. Una vez obtenida esta estructura, convertimos cada nodo a una representación vectorial utilizando el algoritmo Node2Vec explicado en el capítulo anterior, con los parámetros expresados en la tabla 5.3.

Parámetro	**Valor**
Caminatas aleatorias	5
Iteraciones en el procedimiento de aprendizaje	100
Dimensiones de los vectores resultantes	500

Tabla 5.3: Parámetros utilizados para representaciones Node2Vec

En cuanto a las representaciones distribuidas Word2Vec, construimos los vectores monolingües para español y para náhuatl utilizando como corpus de entrenamiento los textos en español y los textos en náhuatl, respectivamente, del corpus paralelo. Estas representaciones fueron generadas utilizando el enfoque CBOW (*Continous Bags Of Words*) y seleccionamos 600 como la dimensionalidad de los vectores resultantes.

Es importante enfatizar que como parte del diseño de la metodología, y de la experimentación, se probaron diferentes parámetros en cada etapa. Por ejemplo, el número de pares de traducción a tomar en cuenta para formar el grafo, la dimensionalidad de los vectores de tipo Node2Vec y Word2Vec, el enfoque para generar vectores tipo Word2Vec

(Cbow o Skip-gram), entre otros. Sólo se reportan los parámetros finales utilizados que dieron los mejores resultados.

Una vez generados los vectores palabra, es necesario aplicar la transformación lineal que se detalló en la sección 4.4.2. Para esta etapa final se necesita del lexicón semilla inducido previamente y de un conjunto de evaluación que contenga pares correctos de traducción español-náhuatl que sirvan como referencia para evaluar. Desde luego, el conjunto de pares de traducción del lexicón semilla y el conjunto de pares de evaluación deben ser disjuntos para evitar sobreajuste, es decir, queremos que la transformación lineal sea capaz de generalizar la transformación de vectores del español a una representación cercana a su traducción, aunque no haya visto estos pares de vectores antes (durante el entrenamiento).

La tabla 5.4 contiene el tamaño del lexicón semilla y el conjunto de evaluación utilizados para la experimentación. El conjunto de evaluación es el mismo presentado en la sección 5.1 para evaluar el impacto de la segmentación morfológica; sin embargo, originalmente se tenían 150 pares de evaluación, en este entorno experimental el número de pares de evaluación decreció a 140. Decidimos eliminar algunas entradas para las cuales no había un vector Node2Vec asociado, ya sea para la palabra en español o alguno de sus candidatos a traducción en náhuatl. Esta es una consecuencia de nuestro planteamiento para obtener representaciones multilingües a partir del método asociativo y realizar un corte de los candidatos para construir el grafo.

Conjunto de datos	**Tamaño**
Lexicón semilla	496 pares de traducción
Conjunto de evaluación	140 pares de traducción

Tabla 5.4: Tamaño de conjuntos de datos utilizados

Una vez construidas las representaciones vectoriales, se aprende la matriz de transformación o mapeo lineal con ayuda del lexicón semilla y se aplica a cada uno de los vectores palabra en español del conjunto de evaluación. Para obtener los candidatos de traducción se utiliza un enfoque de k vecinos, esto es, por cada vector en español que es transformado se calcula mediante similitud coseno su cercanía a todos los vectores del espacio semántico

náhuatl y se ordenan las palabras (morfos) de náhuatl que resultaron más cercanas.

Con el fin de evidenciar la utilidad de aprender una transformación lineal para mejorar el pareamiento entre las unidades de dos lenguas, en nuestra evaluación contrastamos los siguientes entornos experimentales: por un lado evaluamos la tarea de extracción léxica bilingüe utilizando las representaciones vectoriales de tipo Node2Vec sin aplicar ningún mapeo, es decir, calculamos la similitud coseno entre los vectores de las dos lenguas y evaluamos si los pares de traducción están cercanos en el espacio vectorial (*N2V-NOmap*). Por otro lado, evaluamos la misma tarea, pero ahora aplicando la transformación lineal aprendida con el lexicón semilla y utilizando dos tipos de representación vectorial, vectores Node2Vec y también vectores de tipo Word2Vec (*N2V-map*, *W2V-map*).

En la tabla 5.3 de resultados, se puede observar que nuestros vectores bilingües generados a partir de grafos junto con la combinación lineal (*N2V-map*) obtienen la mejor precisión, considerablemente mayor a los demás casos. Se puede observar también que la configuración experimental de Word2Vec más la transformación lineal (*W2V-map*) obtuvo un mal desempeño en nuestro corpus, esto es consistente con los trabajos recientes que han señalado que esta aproximación, aunque es popular, tiende a tener un desempeño deficiente, especialmente si se carece de corpus enormes para entrenar los vectores de cada lengua. En este sentido nuestra propuesta parece ser efectiva para lidiar con el entorno de bajos recursos pues solo necesitamos de un corpus paralelo pequeño (alineado a nivel oración) para crear representaciones vectoriales útiles para la tarea de extracción léxica bilingüe.

Es interesante notar que las representaciones bilingües Node2Vec por si solas (*N2V-NOmap*) son útiles para detectar traducciones (debido a que fueron inicializadas con los resultados del método asociativo); sin embargo, la extracción léxica bilingüe se vuelve aún más precisa cuando se aplica la transformación lineal entre estas representaciones (*N2V-map*). Esto nos da indicios de que el lexicón semilla inducido, en efecto resultó útil para acercar a los vectores en español a su traducción correspondiente en el espacio de náhuatl.

En el apéndice E se pueden consultar los resultados obtenidos con nuestra mejor con-

	P@1(%)	**P@5(%)**	**P@10(%)**
N2V-NOmap	28.2	64	69.9
N2V-map	**67.1**	**88.6**	**91.4**
W2V-map	2	2	4

Tabla 5.5: Evaluación de extracción léxica bilingüe

figuración experimental N2V-map; se presenta un listado de las palabras contenidas en el conjunto de evaluación, así como los 10 candidatos a traducción que resultaron más cercanos.

5.4. Discusión

Hemos presentado los resultados de las diferentes etapas que intervienen en nuestra metodología de extracción léxica bilingüe español-náhuatl. Nuestra metodología trató de mantenerse no supervisada y con cierta independencia de las lenguas con el fin de que pudiera ser fácilmente extensible a otros pares de lenguas en un futuro. Esto quiere decir que tratamos de prescindir de reglas explícitas o bases de conocimiento especialmente confeccionadas para una lengua.

Una de nuestras preguntas de investigación iniciales giraba en torno a si podíamos prescindir de un lexicón semilla pre-compilado para aprender a relacionar los espacios de dos lenguas. Nuestra evaluación sugiere que esto es posible y que nuestro método que combina dos aproximaciones estadísticas para formar un lexicón semilla es de utilidad. Sin embargo, hay ciertas partes de nuestra metodología que podrían considerarse con cierto grado de supervisión o conocimiento específico de las lenguas tratadas. Por ejemplo, nuestra propuesta de inducción de lexicón semilla, así como la de construcción de representaciones vectoriales bilingües, se inicializan con base en resultados de métodos estadísticos de asociación o estimativos para poder relacionar palabras de las dos lenguas; esto es solo posible gracias a que tenemos un corpus paralelo alineado a nivel oración. En este sentido, la alineación a nivel oración puede verse como un proceso supervisado, aunque es importante recordar que la propia alineación a nivel oración puede hacerse con métodos completamente no supervisados (en nuestro caso fue híbrida).

Otro aspecto con cierta dependencia de la lengua es el tratamiento morfológico de los textos, las herramientas utilizadas para el español se basan en reglas o diccionarios, en el caso de la segmentación morfológica nuestros modelos fueron debilmente supervisados, abstraen los patrones de segmentación solo analizando el corpus, esto podría parecer prácticamente independiente de la lengua, sin embargo, las elecciones que realizamos en la metodología como decidir segmentar morfológicamente al náhuatl, o comparar lemas del español con morfos del náhuatl, responden a intuición y conocimiento lingüístico del español y el náhuatl, que no necesariamente se extendería exitosamente para otros pares de lenguas.

La primera parte de la evaluación se enfocó en contestar, de manera empírica, nuestras preguntas de investigación relacionadas con el hecho de si el tratamiento morfológico de las lenguas puede facilitar la tarea de encontrar correspondencias léxicas bilingües. En efecto, trabajar con representaciones a nivel morfo, stem o lema, mostró ser de utilidad para facilitar el análisis estadístico tanto de métodos asociativos como estimativos, así como para reducir el problema de dispersión al construir representaciones vectoriales; todo lo anterior resulta especialmente útil en un entorno de bajos recursos digitales.

Asimismo, se comprobó que la reducción de la complejidad morfológica de las representaciones del texto, en general, está asociada a una mayor precisión en la tarea de extracción léxica bilingüe. Nuestra propuesta de cuantificar la complejidad morfológica usando la distribución de tipos y tokens en el corpus (TTR), arrojó que las representaciones de menor complejidad y más cercanas entre ellas, se obtienen cuando el español y el náhuatl se encuentran segmentados morfológicamente. Sin embargo, el impacto más positivo en nuestra tarea de extracción léxica bilingüe parece ser provocado por las representaciones donde el español está lematizado y el náhuatl segmentado morfológicamente (este es el segundo tipo de representación con menor diferencia de TTRs, tabla 4.9). Decidimos trabajar finalmente con estas representaciones, $ES_{lemma} - NA_{morph}$, no solo por su buen desempeño, sino por la conveniencia de que son pares más cercanos a una entrada de diccionario, esto es, idealmente un lema del español pareado con una forma del náhuatl que no tenga marcas flexivas. Lo anterior provocó que la evaluación de

estos pares fuera más fácil, además, de que estas representaciones podrían permitir de manera más directa automatizar la evaluación usando diccionarios digitales, en un futuro.

En relación con el tipo de evaluación, nuestra evaluación de tipo manual implica varias limitaciones. En primer lugar, el hecho de que uno o más jueces humanos, conocedores de las lenguas, tengan que determinar las posibles traducciones correctas en náhuatl para una palabra en español es una tarea costosa en términos de tiempo o recursos humanos. En nuestro caso esto influyó en que nuestros conjuntos de evaluación tuvieran máximo 150 entradas. En particular tuvimos un sólo juez, la sustentante de esta tesis, que determinó los posibles candidatos de traducción correctos para un conjunto de palabras en español extraídas aleatoriamente del corpus paralelo. Se comentaron anteriormente (sección 5.1) los criterios utilizados para sustentar la rigurosidad de nuestros lexicones de prueba; sin embargo, estamos conscientes de las limitantes de esta evaluación. Es por eso que el conjunto de evaluación construido en este trabajo es de libre consulta en la web[2] y se incluye también en esta tesis (apéndice E) con el fin de que pueda ser evaluado colaborativamente, mejorado y extendido por más especialistas en náhuatl; y pueda volverse referencia para evaluar trabajos posteriores de extracción léxica bilingüe. Se incluye, por cada palabra en español, las posibles traducciones consideradas como correctas por el anotador, así como las traducciones arrojadas por nuestro método automático. En ese listado se pueden analizar los casos en que el método automático extrajo traducciones correctas, de acuerdo a las traducciones de referencia, y los casos en que no fue posible extraer una traducción correcta en los primeros diez candidatos.

En cuanto a la flexibilidad de nuestra evaluación manual, vale la pena comentar algunos aspectos. Es importante tener en cuenta que la naturaleza de nuestro corpus paralelo, a pesar de ser una selección de textos más o menos regulares, sigue siendo diversa, con variación ortográfica y dialectal, así como errores de OCR no detectados. Por otro lado, nuestra segmentación morfológica del náhuatl es automática y semisupervisada, sujeta a errores de segmentación y desde luego con resultados que carecen, hasta cierto punto, de la fineza de análisis hecha por un lingüista. Todo lo anterior contribuye, en mayor o menos medida, a que los candidatos de traducción obtenidos por nuestros métodos no siempre

[2]https://sites.google.com/site/xgutierrezv/lexicon-espanol-nahuatl-automatico

tengan la forma ideal de una entrada de diccionario. Esto nos obligó a ser flexibles con los problemas de segmentación morfológica, es decir, tomamos como correcto un par de traducción, aunque su parte en náhuatl tuviera un error de segmentación.

Podemos ver el ejemplo en la tabla 5.6 de la entrada en español "comer", algunos opinarían que su traducción correcta sería "cua", que es la forma no marcada de comer (3SG.PRS). Sin embargo, en nuestra evaluación también admitimos como posibilidades correctas a formas como "cuaz", "cualtiz" y otras formas del verbo que aparecen dentro de los candidatos, aunque tengan aglutinadas marcas de potencialidad o causatividad por un presunto error del segmentador automático.

Otro tipo de conflicto durante la evaluación fueron los préstamos, es decir, palabras que no cambian entre las lenguas. Lo que es relativamente común encontrar en nuestro corpus, de manera particular, fue problemático evaluar aquellas préstamos del español al náhuatl, pues cuando el modelo morfológico del náhuatl se enfrentó a estas formas en español, realizó segmentaciones, desde luego, incorrectas. En estos casos decidimos también ser flexibles y evaluarlos como correctos si el error había sido de segmentación. En la tabla 5.6 se puede ver el ejemplo de la palabra "centenario": probablemente en el corpus paralelo esta palabra no cambia traducida al náhuatl, por lo que el segmentador morfológico al encontrarlo lo segmenta como *-nario* y este morfo resulta el primer candidato de traducción. En este tipo de casos, tomamos la traducción como correcta.

En relación con la efectividad de las diferentes representaciones vectoriales, resulta destacable la precisión de nuestros vectores bilingües Node2Vec mapeados con la transformación lineal. Aunque este no es el primer trabajo que combina estructuras de grafos y representaciones vectoriales, métodos similares son probados exclusivamente en entornos de gran cantidad de corpus disponibles para construir las representaciones (Pelevina et al., 2017; Newman-Griffis y Fosler-Lussier, 2017; Blunsom y Hermann, 2014; Kočiskỳ et al., 2014). Nuestros experimentos no solo muestran una mejora significativa en comparación con solo usar vectores monolingües Word2Vec, sino que las precisiones obtenidas con Word2Vec son tan bajas que se podría concluir que este escenario experimental, aunque popular en PLN (Mikolov et al., 2013b), no tiene una aplicación práctica cuan-

Palabra_ES: comer Traduccion_NA: cua:0.815 cuaz:0.814 tos:0.731 tlacual:0.721 cualtiz:0.66 tepahpaquiltia:0.607 zque:0.586 dral:0.550 cuap:0.549 niquintlayocoli:0.546 Referencia: ['cuaz', 'cualtiz', 'otlacua', 'cua', 'tlacual'] Rank: 1
Palabra_ES: centenario Traducción_NA: nario:0.914 premios:0.821 icuilacta:0.795 ipanoque:0.739 presi:0.695 tazohta:0.649 hcuilo:0.601 elehui:0.597 diego:0.590 dral:0.586 Referencia: ['nario'] Rank: 1

Tabla 5.6: Ejemplos de candidatos a traducción problemáticos

do se tienen corpus pequeños para entrenar a las representaciones vectoriales de este tipo.

Las representaciones distribuidas tipo Word2vec codifican el significado de las palabras por medio de sus contextos de aparición, es por eso que una de nuestras preguntas iniciales de investigación era si podríamos extraer pares de traducción comparando, de alguna manera, los contextos que modelan a las palabras de cada lengua. Sin embargo, la evaluación fue de utilidad para reflexionar que mientras este planteamiento es coherente y tiene fundamentos lingüísticos, las representaciones vectoriales populares de hoy en día en PLN aún no alcanzan a capturar de manera exitosa el significado cuando las palabras aparecen pocas veces en un corpus y, por lo tanto, no hay suficientes contextos de aparición. Es importante mencionar que, aunque no aparece reportado en la tesis, durante toda la experimentación se hicieron pruebas con diversos hiperparámetros y tipos de representaciones, entre ellas, vectores de espacios semánticos distribucionales que no representaron una mejora ni un escenario diferente al que se comenta.

En vista de lo anterior, propusimos un enfoque para mejorar las representaciones vectoriales, cuya efectividad se basa en tener una distribución *prior* de pares de traducción (obtenida a partir de asociaciones estadísticas) que sirve para construir el grafo y posteriormente obtener vectores que codifiquen esta estructura. Una de las limitaciones que encontramos con este enfoque fue que algunas palabras del corpus no obtuvieron una representación vectorial con nuestra metodología. La razón de esto subyace en el método asociativo anymalign y en el umbral de score que establecimos para construir el grafo. Anymalign, al no ser un método de tipo estimativo, no aproxima una distribución de probabilidad que pueda estimar la probabilidad de traducción de cualquier par de palabras en el corpus. Más bien, estima la asociación entre un par de palabras basado en cuántos pequeños subcorpus (submuestreo) estas palabras compartieron la misma distribución. Lo anterior puede provocar que existan pares de palabras para las cuáles nunca se calcule una asociación o que resulte muy baja; esto está en función del tiempo en que se deja correr al método, es decir, del número de submuestreos que alcanza a hacer y, también, de la naturaleza y tamaño del corpus paralelo.

A pesar de lo anterior, el método Anymalign mostró consistentemente mejor desempeño que el método estimativo (modelo IBM-1). De la misma manera, nuestra propuesta de

representaciones bilingües inicializadas con este método se desempeñan muy por encima de Word2Vec a la hora de aprender un mapeo lineal entre las lenguas. Por lo tanto, se podría decir que nuestro método puede perder un poco de cobertura en el número de palabras del corpus que alcanza a modelar, sin embargo su ganancia en precisión es significativa.

Finalmente, en términos absolutos la precisión obtenida en la extracción léxica como resultado de toda nuestra metodología es de 67.1 % (p@1). Esta proporción de pares de traducidos correctamente podría parecer baja, en el sentido de que no es cercana al 100 %; sin embargo, es importante tomar como referencia los métodos del estado del arte similares que inducen pares de traducción mediante representaciones vectoriales y un mapeo, estos tienen desempeños, en el mejor de los casos, de alrededor del 45 % para algunos pares de lenguas (Artetxe et al., 2018). Desde luego estos métodos tienen una formulación un poco distinta a la nuestra, pues parten de representaciones como Word2Vec obtenidas a partir de grandes corpus monolingües sin necesidad de tener un corpus paralelo. Sin embargo, nuestra evidencia muestra que es suficiente un corpus paralelo pequeño para alcanzar resultados equiparables o mejores. Lo anterior es consistente con trabajos recientes como Levy et al. (2016), donde demuestran que agregar una señal "supervisada", por ejemplo un conjunto relativamente pequeño de oraciones paralelas, mejora considerablemente las tareas que utilizan representaciones vectoriales croslingües extraídas de corpus gigantescos monolingües y que no son capaces de superar un umbral de desempeño.

Por otro lado, si contrastamos los resultados de nuestro método (tabla 5.3) con los resultados del método asociativo Anymalign (cuando fue probado por si solo para evaluar el impacto de la morfología, tabla 5.1) podemos notar que la diferencia de precisión no es muy grande, lo cual podría cuestionarnos sobre el alcance de nuestra contribución. Sin embargo, un aspecto que es importante tener en cuenta es que nuestro método no solo incluye la etapa de extracción léxica bilingüe, sino la creación de representaciones vectoriales bilingües. El hecho de haber construido representaciones vectoriales nos da la flexibilidad de ocupar en un futuro estas representaciones del español y del náhuatl para una gran gama de aplicaciones de PLN que hacen uso de vectores, por ejemplo, realizar operaciones composicionales para construir representaciones vectoriales más completas

del significado (Baroni et al., 2014), o aplicar métodos de aprendizaje de máquina sobre los vectores obtenidos.

Tomando en cuenta el análisis de los resultados obtenidos, y a manera de recapitulación final, el esquema de la figura 5.2 resume el procedimiento que derivó en los mejores resultados para la extracción léxica bilingüe español-náhuatl.

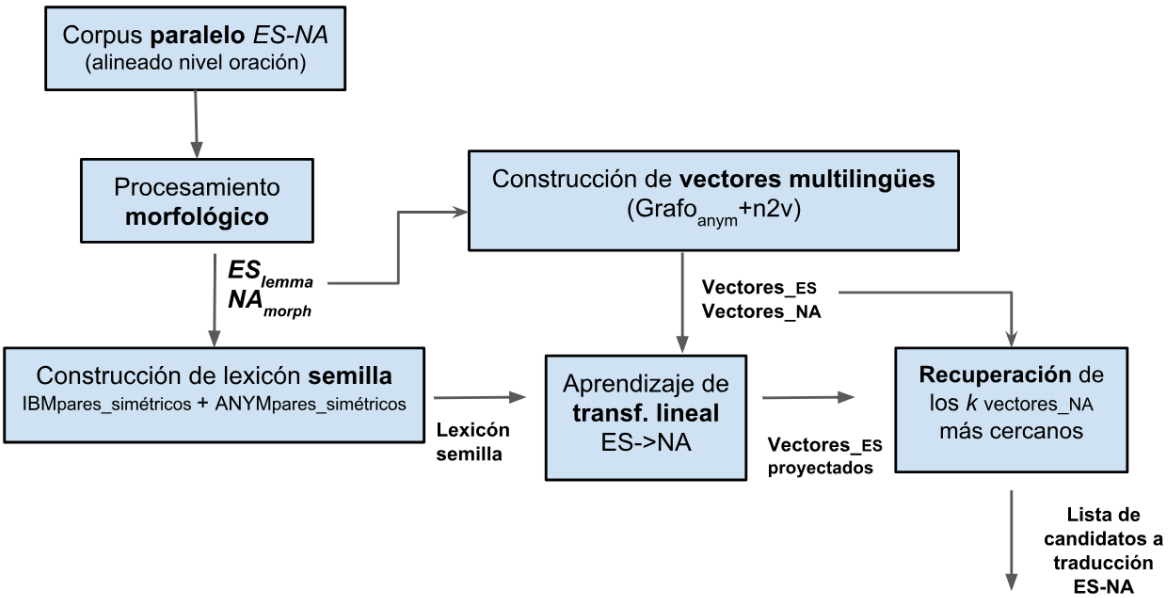

Figura 5.2: Esquema final de extracción léxica bilingüe español-náhuatl

Capítulo 6

Conclusiones

Este trabajo gira en torno a la investigación y desarrollo de tecnologías del lenguaje para lenguas de bajos recursos digitales. Específicamente nos hemos centrado en la extracción automática de léxico bilingüe para el par de lenguas español-náhuatl. Esta es una tarea que implica diversos retos que fueron analizados a lo largo de la tesis. En primer lugar, la escasez de corpus es una gran limitante, pues no permite aplicar de manera exitosa los métodos populares de PLN que suelen basarse en colecciones muy grandes de documentos disponibles para una lengua. En segundo lugar, las diferencias tipológicas lingüísticas, principalmente a nivel morfológico, entre el español y el náhuatl aumentan la complejidad en la alineación o extracción de correspondencias bilingües a nivel de palabra. Finalmente, nos enfrentamos también a una serie de retos metodológicos y/o tecnológicos derivados de características intrínsecas del náhuatl, esto es, su falta de normalización ortográfica, su gran variación dialectal, escasez de recursos para preprocesar esta lengua, entre otros.

Con base en lo anterior, nuestras hipótesis para resolver el problema implicaron una reformulación o adaptación de las aproximaciones estándar de extracción léxica bilingüe. La metodología presentada en esta tesis está destinada a entornos en donde se tiene un corpus paralelo de tamaño pequeño. Es importante destacar nuestra intención de diseñar una metodología con enfoques lo menos supervisados posibles para no depender de recursos externos, difíciles de conseguir principalmente para el náhuatl. Esto no implica que prescindimos de conocimiento e intuición lingüística, al contrario, el conocimiento general de los fenómenos del español y el náhuatl fue muy importante para discernir las soluciones apropiadas.

Nuestro trabajo involucró tomar en cuenta la morfología de las lenguas para lograr representaciones que permitieran una mejor alineación de correspondencias léxicas bilingües. También, normalizamos la variación morfológica como una estrategia para combatir el problema de unidades léxicas con baja frecuencia o de dispersión de los datos, un fenómeno especialmente problemático cuando se procesan lenguas morfológicamente ricas con poco corpus disponible.

Comprobamos de manera cuantitativa que el náhuatl, una lengua aglutinante con tendencia polisintética, se beneficia del uso de herramientas de segmentación morfológica que ayudan a reducir ampliamente la complejidad morfológica de los textos. Para el español, una lengua primordialmente fusional y sufijal en su flexión, es suficiente con aplicar stemming o lematización para reducir la variación morfológica. Posteriormente, comprobamos que los distintos métodos de extracción léxica bilingüe mejoraron sus resultados cuando el corpus paralelo fue procesado morfológicamente. En particular, obtuvimos el mejor desempeño al parear correspondencias entre lemas del español y morfos del náhuatl. Todo esto resulta congruente con las tendencias recientes en PLN que buscan diseñar modelos y representaciones a nivel subpalabra para mejorar una gran diversidad de tareas.

Por otro lado, nuestra estrategia de extracción de léxico bilingüe consistió en la combinación de diferentes métodos estadísticos y características para estimar los pares de traducción. El objetivo era poder llegar, en última instancia, a representaciones vectoriales de las dos lenguas que permitieran extraer pares de traducción a partir de la distancia o similitud que comparten estos vectores. Esto último representa un planteamiento muy popular hoy en día y suele implementarse con vectores distribuidos (Word2Vec principalmente) que capturan la semántica distribucional de las palabras, así como con una transformación o mapeo que permita relacionar los espacios vectoriales de dos lenguas.

Tuvimos que adaptar este enfoque para que funcionara en un entorno experimental como el nuestro. Nuestro lexicón semilla, utilizado para aprender una transformación lineal que mapea los vectores de español a náhuatl, fue inducido de manera prácticamente no supervisada (sólo se necesitó del corpus paralelo alineado a nivel oración). La inducción de este lexicón semilla se logró combinando un enfoque asociativo para extraer pares de

traducción basado en sub-muestreo del corpus (*ANYM*) , así como un enfoque estimativo para aprender distribuciones de probabilidad de traducción léxica (*IBM-1*).

Asimismo, las representaciones vectoriales distribuidas Word2Vec, altamente populares en la literatura de PLN, resultaron no ser de utilidad para representar las palabras, cuando fueron generadas con nuestro corpus paralelo (se necesitan corpus de mayor dimensión). Debido a esto, propusimos un tipo de representación vectorial multilingüe que parte de la idea de generar un grafo que relacione las palabras de las dos lenguas: utilizando scores obtenidos del método basado en sub-muestreo del corpus (*ANYM*) y a partir de este grafo construir representaciones vectoriales densas (Node2Vec). Usando estas representaciones, de manera natural los vectores de la lengua fuente están más cercanos a su traducción correspondiente en la lengua destino (por la distribución a priori contenida en el grafo). Cuando aplicamos la transformación lineal entre estos vectores, fuimos capaces de acercar aún más los vectores de la lengua fuente a su traducción en la lengua destino.

6.1. Aportaciones científicas

Entre las contribuciones derivadas de esta tesis se pueden destacar las siguientes: (i) la constitución, digitalización y puesta en línea de un corpus paralelo español-náhuatl; (ii) la generación de modelos semisupervisados de segmentación morfológica tanto para el español como para el náhuatl; (iii) una metodología para cuantificar la variación de complejidad morfológica cuando diferentes tipos de preprocesamiento morfológico son aplicados a una lengua; (iv) una metodología híbrida, basada en métodos de tipo estadístico, para estimar correspondencias léxicas bilingües español-náhuatl a partir de un corpus paralelo; (v) un método para obtener representaciones vectoriales bilingües; (vi) un lexicón bilingüe español-náhuatl generado de manera automática, sin intervención de un humano.

El apéndice (ref) contiene un listado de publicaciones científicas que fueron resultado de algunos de los puntos previamente mencionados.

Quizá el resultado práctico más evidente de esta tesis sea la generación de un lexicón

bilingüe español-náhuatl. Esto no pretende sustituir la labor manual de un especialista de estas lenguas. Sin embargo, es un recurso digital útil que puede facilitar tareas e incluso alimentar a sistemas de alto nivel de PLN, por ejemplo, de traducción automática español-náhuatl. Los rankings obtenidos de correspondencias léxicas bilingües pueden sustituir o complementar las tablas de traducción, un componente de los modelos estadísticos de traducción, y mejorar así la traducción automática para este par de lenguas, incluso la de tipo neuronal (He et al., 2016; Pourdamghani et al., 2018).

Esperamos que nuestro trabajo pueda ser de utilidad para aplicarse a pares de lenguas que enfrentan condiciones similares; contribuyendo así al desarrollo del PLN en entornos de bajos recursos digitales. Hoy en día los escenarios de escasos recursos, o pocos datos, se han vuelto un problema de investigación que ha atraído la atención de áreas como el PLN y el aprendizaje de máquina. Avances importantes se esperan en el futuro en este ámbito, lo cual tendrá un impacto positivo en el mejoramiento y creación de tecnología para muchas lenguas del mundo.

Además del impacto en términos de generación de tecnologías del lenguaje, consideramos que este tipo de trabajos son propicios para incentivar la discusión lingüística sobre los alcances y limitaciones de los enfoques de PLN cuando son aplicados a lenguas con características como las que enfrentamos. Así, la investigación en PLN, o lingüística computacional, puede ser una herramienta que ayude a profundizar en el entendimiento del lenguaje humano, sus estructuras e, idealmente, contribuir a construir modelos más generales de la lengua.

Finalmente, es importante reiterar la gran diversidad y riqueza lingüística que posee México, convirtiendo a este país en un escenario ideal para la investigación y desarrollo de modelos computacionales capaces de lidiar con estas lenguas. Lo anterior no sólo representa un interesante reto desde la perspectiva científica, sino, una oportunidad para contribuir en la generación de recursos y tecnologías destinadas a los hablantes y estudiosos de nuestras lenguas originarias.

6.2. Trabajo futuro

Existen aún varias direcciones de investigación que se pueden explorar dentro de la tesis. En esta sección presentamos un breve panorama de estas posibles extensiones.

Constitución del corpus paralelo digital

La compilación y digitalización del corpus paralelo implicó diversos retos tecnológicos y metodológicos que se expusieron a lo largo de esta tesis. La tarea de corrección de los documentos digitalizados fue primordialmente manual, lo cual resultó costoso y difícilmente extensible para nuevos textos. En un futuro, se podría hacer uso de trabajos recientes de PLN que implementan sistemas de OCR capaces de lidiar con la variación ortográfica de textos históricos, incluido el náhuatl (Garrette y Alpert-Abrams, 2016).

La gran diversidad dialectal y ortográfica presente en los documentos en náhuatl, nos obligó a tomar la decisión metodológica de trabajar solo con un subconjunto del corpus. Sin embargo, se podría profundizar en estrategias para normalizar esta diversidad y poder trabajar así con una proporción más grande de textos del corpus. Esto tendría un impacto positivo en el desempeño de los métodos estadísticos de PLN.

En el futuro nos gustaría agregar mayor información al corpus en línea, esto es, anotar variante dialectal, norma ortográfica, periodo, etc. Lo anterior se traduciría en un mayor número de metadatos, en el sistema de recuperación de información, que permitirían búsquedas y estudios más completos a partir del corpus paralelo en línea.

Vale la pena recordar que el corpus paralelo de esta tesis constituye también un proyecto web de consulta pública (Axolotl), que posee un grupo creciente de usuarios, por lo que resulta importante mantener un continuo enriquecimiento de fuentes paralelas con el fin de incrementar el tamaño del corpus en línea español-náhuatl. Fomentando así el desarrollo de estudios y tecnologías para este par de lenguas.

Tratamiento morfológico de las lenguas

En esta tesis aplicamos enfoques semisupervisados, basados en teoría de la infomación, para realizar la segmentacíón morfológica del náhuatl. Los resultados obtenidos fueron lo suficientemente útiles para el desarrollo de nuestra metodología; sin embargo, también es posible observar que las segmentaciones obtenidas son perfectibles. Nuestra intención original era combinar, de alguna manera, tanto la herramienta basada en reglas (Chachalaca) como la alternativa semisupervisada; de tal manera que pudiéramos estimar una segmentación morfológica más robusta, es decir, que combinara las fortalezas de los métodos estadísticos y de la rigurosidad lingüística.

La integración del software Chachalaca (Thouvenot, 2011) a nuestra línea de experimentos fue problemática y con costo computacional prohibitivo. Sin embargo, nos gustaría resaltar que el analizador morfológico Chachalaca constituye un recurso valioso para el análisis automático de náhuatl y que tiene un interesante potencial de ser reimplementado para facilitar su portabilidad y uso. Por ejemplo, se podrían utilizar máquinas de estados finitos (Hulden, 2009).

Por otra parte, en la actualidad modelos neuronales de tipo secuencia a secuencia (seq2seq) han probado ser útiles en la tarea de segmentación morfológica, incluso aplicados a lenguas de bajos recursos con morfología aglutinante polisintética (Kann et al., 2018). Esta constituye una interesante alternativa para la segmentación morfológica del náhuatl.

En cuanto análisis más del tipo lingüístico, se podría hacer una valoración cuantitativa y cualitativa de los patrones de segmentación morfológicos obtenidos con los modelos entrenados en esta tesis. Esos patrones podrían revelar interesantes fenómenos de la morfología del náhuatl o del español. Podrían, también, ser una herramienta para detectar las limitaciones de los métodos no supervisados de segmentación morfológica y mejorarlos.

Finalmente, en esta tesis exploramos el concepto de complejidad morfológica como una manera de cuantificar qué representaciones de los textos son menos complejas y, por lo tanto, más fáciles de alinear entre lenguas. De manera más general, el estudio de la complejidad lingüística constituye un reto muy interesante con implicaciones no sólo en el

estudio de fenómenos tipológicos o de adquisición de la lengua, sino en PLN. Recientemente, en nuestra área, han surgido trabajos que resaltan el impacto de cuantificar la complejidad de las lenguas, como una forma para determinar qué modelos de PLN pueden extenderse exitosamente a un número grande de lenguas y qué otros no (Bentz et al., 2016; Cotterell et al., 2018).

En este sentido, el estudio cuantitativo de la complejidad morfológica de las lenguas mexicanas constituye una importante línea de investigación futura.

Extracción léxica bilingüe

En cuanto al mejoramiento de nuestra metodología para extracción léxica bilingüe, existen varias direcciones posibles. De manera general, nuestro planteamiento es flexible para reemplazar los esquemas de pesado, o score, que utilizamos tanto para inducir el lexicón semilla como para generar representaciones vectoriales bilingües. Es decir, el método asociativo basado en sub-muestreo (*ANYM* y el método estimativo (*IBM-1*) pueden sustituirse por otro tipo de métodos sin afectar la arquitectura general de nuestra metodología.

En particular, sería interesante probar esquemas más sofisticados para inicializar el grafo que relaciona las palabras entre español y náhuatl. En esta tesis probamos diversas variaciones entre el número de candidatos de traducción y scores para inicializar el grafo; finalmente sólo usamos scores derivados del método asociativo (*ANYM*) . Sin embargo, queda mucho espacio para mejorar la construcción de este grafo bilingüe y que, por tanto, derivará en la obtención de mejores representaciones vectoriales bilingües del tipo Node2Vec. Una alternativa interesante es el uso de redes convolucionales de grafos (Graph Convolutional Networks) para producir representaciones vectoriales a partir de grafos (Bruna et al., 2013), así como para realizar tareas de aprendizaje de máquina aplicadas a la extracción bilingüe.

La inducción de pares de traducción por medio de una transformación lineal, utilizando corpus paralelos, comparables o monolingües, constituye un tema muy popular en PLN hoy en día. Una gran diversidad de variaciones se han propuesto para enriquecer este

planteamiento. Por nuestra parte, creemos que sería interesante tratar de aprender una transformación de tipo no lineal que relacione los espacios entre dos lenguas. Lo anterior se podría lograr con arquitecturas profundas de aprendizaje neuronal.

También se deben mencionar las alternativas de transferencia del conocimiento hacia lenguas de bajos recursos digitales. Esto engloba un conjunto de enfoques que se enfrentan a lenguas objetivo que carecen de recursos anotados, sin embargo, son capaces de transferir conocimiento de lenguas de mayores recursos. Por ejemplo, estos enfoques se han utilizado para transferencia de etiquetado POS, análisis sintáctico, etiquetado de roles semánticos. Este camino es también prometedor para el desarrollo de tecnologías del lenguaje para lenguas como el náhuatl.

Finalmente, el procesamiento automático de las lenguas de bajos recursos también se puede beneficiar de los avances teóricos en el área de aprendizaje de máquina que buscan lidiar mejor con la escasez de datos de entrenamiento

Evaluación

Nuestra metodología está formada por diferentes etapas, varias de estas fases fueron evaluadas con un esquema en donde intervino el conocimiento de un humano para anotar conjuntos de referencia de evaluación, particularmente en los experimentos de segmentación morfológica y de extracción léxica bilingüe. En PLN, es usual que en las evaluaciones donde intervienen juicios humanos se contemple la participación de más de un juez y el uso de medidas de acuerdo y de sesgo para asegurar la confiabilidad de la evaluación. En el capítulo de evaluación y análisis de resultados (5) discutimos porqué este procedimiento estándar no es fácilmente extrapolable a entornos de bajos recursos como el nuestro. En el futuro nos gustaría incluir más juicios humanos para fortalecer nuestros conjuntos de evaluación. Por ahora el conjunto de evaluación de extracción léxica bilingüe, desarrollado como parte de esta tesis, es público y abierto a recibir retroalimentación de especialistas en náhuatl. Los conjuntos de evaluación de segmentación morfológica del español y el náhuatl pertenecen a sus autores.

Finalmente, gran parte de las limitaciones están relacionadas con el hecho de lo cos-

toso que resulta una evaluación manual. En un futuro, nuestros conjuntos de evaluación podrían hacerse más grandes con la ayuda de métodos semiautomáticos que extraigan pares de traducción de referencia de algunas fuentes externas, por ejemplo, diccionarios digitales. Sin embargo, la tarea no es trivial debido a la variación ortográfica; tambíen, debido al hecho de que nuestro lexicón bilingüe contiene lemas del español alineados a raíces o radicales del náhuatl y no siempre los diccionarios español-náhuatl contienen estas formas.

Otro conflicto, que encontramos durante la evaluación, fueron los errores de segmentación morfológica. En este trabajo decidimos adoptar un criterio laxo y tomar como correcto un par de traducción que tuviese un error en la segmentación en la parte náhuatl. Esto perjudica la calidad de lexicón bilingüe extraído, por ejemplo, si se quisiera construir un diccionario español-náhuatl a partir de este lexicón, se necesitaría de una revisión manual por parte de un especialista para verificar que los pares extraído tengan el formato esperable de una entrada de diccionario bilingüe. En el futuro, podría diseñarse una evaluación que incorpore la dimensión de los errores causados por la segmentación morfológica.

Apéndice A

Documentos del corpus paralelo

Título	**Variante náhuatl**
Adivinanzas nahuas de hoy y siempre	Actual
Amerindia, leyendas y cantos nahuas.	Actual
Anales de Tepeteopan, De Xochitecuhtli a don Juan de San Juan Olhuatecatl	Clásico
Antología del cuento náhuatl	Actual
Augurios y abusiones	Clásico
Cantos indígenas de México	Actual
Chimalpain Cuauhtlehuanitzin.Primera, segunda, cuarta, quinta y sexta relaciones de las differéntes histories originales	Clásico
Documentos nauas de la Ciudad de México del siglo XVI	Clásico
El anillo de Tlalocan	Actual
El náhuatl de Tetzcoco en la actualidad	Actual
Historia de México narrada en náhuatl y español de acuerdo al calendario azteca	Clásico
La llave del náhuatl	Clásico
La tierra nos escucha	Actual
La tinta negra y roja, antología de poesía náhuatl	Clásico
La voz profunda Antología de literatura mexicana en lenguas indígenas	Actual
Lo que relatan de antes. Cuentos tének y nahuas de la Huasteca	Actual
Los cuentos en náhuatl de Doña Luz Jiménez	Actual
Método autodidáctico náhuatl-español	Actual
Mitos y cuentos nahuas de la Sierra Madre Occidental	Actual

Título	**Variante náhuatl**
Nican Mopohua	Clásico
Recetario Nahua de Milpa Alta	Actual
Recetario Nahua del norte de Veracruz	Actual
Teatro Náhuatl	Clásico
Testimonios de la antigua palabra	Clásico
Trece poetas del Mundo Azteca	Clásico
Veinte himnos sacros de los nahuas	Clásico
Vida económica de Tenochtitlan	Clásico
Yancuitlalpan. Tradición y discurso ritual	Actual
Revista Lengua y Cultura Náhuatl	Actual
Textos encontrados en internet:	
De Porfirio Díaz a Zapata, memoria náhuatl de Milpa Alta	Actual
Constitución política de los Estados Unidos Mexicanos	Actual
Las ocho relaciones y el memorial de colhuacan	Clásico
Tlahtlapowaltin, cuentos	Actual

Apéndice B

Estructura morfológica del náhuatl

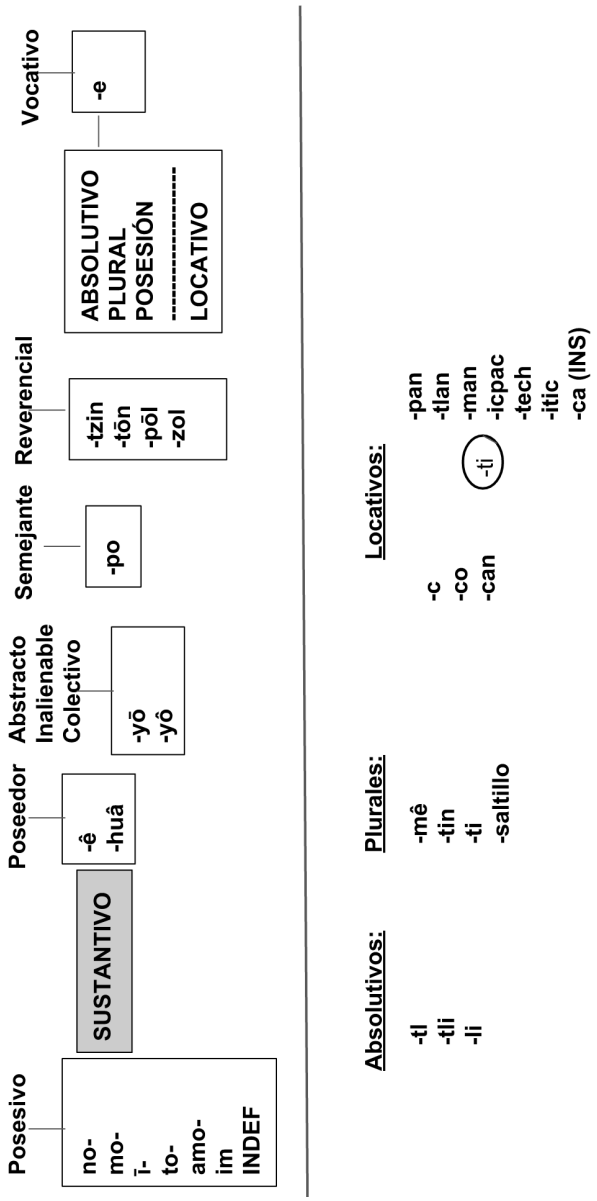

Figura B.1: Estructura de la palabra nominal (basado en el curso de Náhuatl Clásico de Leopoldo Valiñas Coalla)

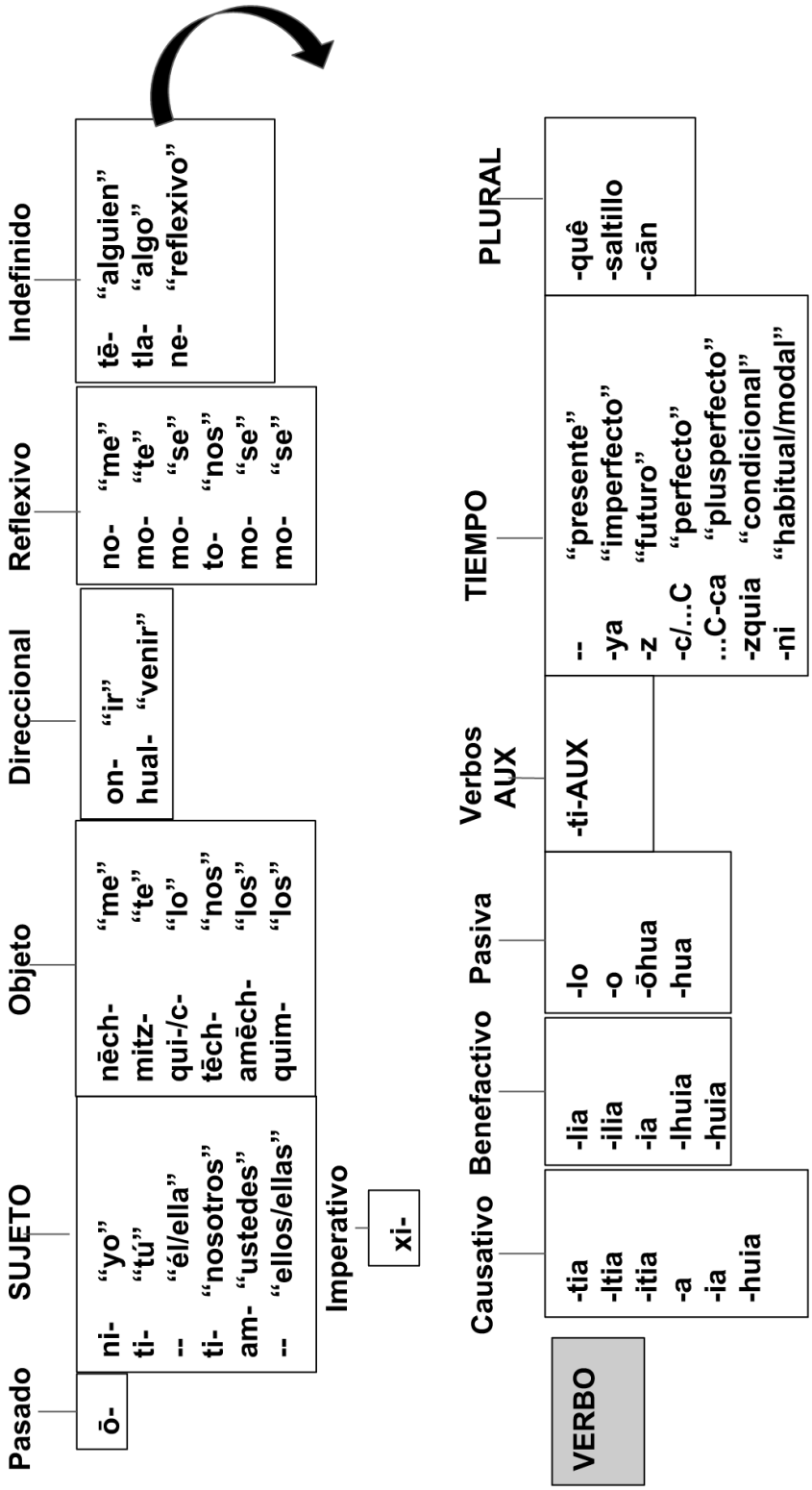

Figura B.2: Estructura de la palabra verbal (basado en el curso de Náhuatl Clásico de Leopoldo Valiñas Coalla)

Apéndice C

Parámetros de segmentación morfológica

- NÁHUATL

Conteo de palabras: Normal (frecuencias)			
$\boldsymbol{\alpha}$	**Pre**	**Rec**	**F**
0.1	67.7 %	89.3 %	77 %
0.2	69.6 %	87.7 %	**77.6** %
0.3	70.1 %	84.9 %	76.8 %
0.4	70.9 %	83.6 %	76.7 %
0.5	71.2 %	82.3 %	76.4 %
0.6	72.5 %	81.6 %	76.8 %
0.7	73.5 %	79.6 %	76.4 %
0.8	75.1 %	80 %	77.5 %
0.9	75.5 %	77.4 %	76.4 %
1	76.6 %	77.4 %	77 %
3	84.5 %	49.6 %	62.5 %
10	97.6 %	22.9 %	37.1 %
-	76 %	75.7 %	75.9 %

Conteo de palabras: Sólo tipos

α	**Pre**	**Rec**	**F**
0.1	67.9 %	89.1 %	77.1 %
0.2	69.4 %	86.8 %	77.1 %
0.3	70.1 %	85.3 %	76.9 %
0.4	71.1 %	83.7 %	76.9 %
0.5	71.8 %	82.9 %	76.9 %
0.6	72.9 %	81.2 %	76.8 %
0.7	74 %	81.3 %	77.5 %
0.8	73.8 %	78.7 %	76.2 %
0.9	75.1 %	76.6 %	75.9 %
1	75.5 %	76.8 %	76.1 %
3	85.9 %	51.6 %	64.5 %
10	96.9 %	22.3 %	36.3 %
-	76 %	76.6 %	76.3 %

- ESPAÑOL

Conteo de palabras: Logaritmo de la frecuencia

α	**Pre**	**Rec**	**F**
0.1	68.2 %	88.1 %	76.9 %
0.2	69.2 %	86.3 %	76.8 %
0.3	70.8 %	83.7 %	76.7 %
0.4	73.2 %	82.2 %	77.4 %
0.5	74.5 %	79.1 %	76.7 %
0.6	75 %	76.7 %	75.9 %
0.7	76.2 %	72.3 %	74.2 %
0.8	80.1 %	69.6 %	74.4 %
0.9	81.6 %	67.4 %	73.8 %
1	81.2 %	62.9 %	70.9 %
3	88 %	43.7 %	58.4 %
10	96.5 %	24.9 %	39.5 %
-	82.6 %	66.2 %	73.5 %

Conteo de palabras: Normal (frecuencias)			
$\boldsymbol{\alpha}$	**Pre**	**Rec**	**F**
0.1	71.8 %	77.7 %	74.6 %
0.2	75.6 %	78.5 %	77 %
0.3	77.1 %	77.7 %	77.4 %
0.4	78.9 %	77.8 %	78.3 %
0.5	78.6 %	76.4 %	77.5 %
0.6	79.6 %	75.8 %	77.7 %
0.7	81.2 %	76.2 %	78.6 %
0.8	83.8 %	77.2 %	80.3 %
0.9	82.9 %	74.9 %	78.7 %
1	83.2 %	75.2 %	79 %
3	90.2 %	63.4 %	74.5 %
10	98.5 %	17.3 %	29.4 %
-	84.5 %	74.8 %	79.3 %

Conteo de palabras: Sólo tipos

α	**Pre**	**Rec**	**F**
0.1	72 %	78.4 %	75.1 %
0.2	74.9 %	77.5 %	76.2 %
0.3	77.3 %	78.3 %	77.8 %
0.4	79.1 %	77.5 %	78.3 %
0.5	79.3 %	77.2 %	78.2 %
0.6	82.1 %	78.7 %	80.4 %
0.7	81.1 %	76.8 %	78.9 %
0.8	81.8 %	76 %	78.8 %
0.9	83.4 %	75.2 %	79.1 %
1	82.6 %	74.2 %	78.1 %
3	89.2 %	62.8 %	73.7 %
10	98.4 %	17.4 %	29.6 %
-	82.6 %	75.3 %	78.8 %

Conteo de palabras: Logaritmo de la frecuencia

α	**Pre**	**Rec**	**F**
0.1	77.1 %	82.6 %	79.7 %
0.2	80.8 %	82.1 %	81.4 %
0.3	82.2 %	81.6 %	81.9 %
0.4	84 %	80.6 %	**82.3** %
0.5	83.9 %	78.4 %	81.1 %
0.6	85 %	77.8 %	81.3 %
0.7	86.3 %	77 %	81.4 %
0.8	86 %	75.8 %	80.6 %
0.9	86.4 %	75.5 %	80.6 %
1	87.2 %	75.5 %	80.9 %
3	89.8 %	66.3 %	76.2 %
10	97.9 %	24.4 %	39.1 %
-	86.8 %	74.9 %	80.4 %

Apéndice D

Ranking de morfos más frecuentes en el corpus paralelo (náhuatl)

1	in
2	i
3	mo
4	ti
5	qui
6	tla
7	tl
8	o
9	li
10	huan
11	ye
12	ca
13	c
14	te
15	a
16	ya
17	h
18	z
19	tli
20	tzin
21	on
22	que
23	co
24	pan
25	ma
26	to
27	no
28	pa
29	yo
30	ce
31	ni

Apéndice E

Conjunto de evaluación e impresión de resultados

```
1 Palabra_ES: jadear
Traducción_NA:
tixtomahua:0.935
itlachi:0.755
tolini:0.734
cuitl:0.631
icemel:0.622
timamalihuac:0.621
ili:0.619
ixpetla:0.615
ixeo:0.604
milacatzo:0.565
Referencia: ['tixtomahua']
Rank: 1

2 Palabra_ES: visceras
Traducción_NA:
tlahuicxitilli:0.795
cuacualaca:0.779
itzoltilia:0.707
motlali:0.697
zoqui:0.693
paca:0.686
tlalia:0.683
temolcax:0.681
caxi:0.680
cuacua:0.673
Referencia: ['ihti', 'ehua']
Rank: 52

3 Palabra_ES: desear
Traducción_NA:
elehui:0.925
elehuia:0.728
qu:0.596
znequi:0.586
quechi:0.581
monequilti:0.566
achto:0.554
nequia:0.523
nec:0.521
nequi:0.520
Referencia: ['elehuia', 'znequi', 'elehui', 'nequi']
Rank: 1

4 Palabra_ES: galleta
Traducción_NA:
galleta:0.964
comal:0.582
gr:0.565
yb:0.565
xacualoa:0.562
beethoven:0.560
pohualli:0.559
rafael:0.559
castillo:0.549
acamapich:0.549
Referencia: ['galleta', 'tlaxcal']
Rank: 1

5 Palabra_ES: horno
Traducción_NA:
horno:0.972
tu:0.751
timihtil:0.627
tz:0.539
cuepa:0.518
teoticanantzin:0.509
cuep:0.507
temi:0.507
tlananquili:0.498
huic:0.483
Referencia: ['horno']
Rank: 1

6 Palabra_ES: olvidar
Traducción_NA:
huilli:0.750
quil:0.658
cahua:0.656
mihtalhui:0.627
el:0.614
xquin:0.548
```

```
xonm:0.544
peh:0.543
cuecue:0.525
nino:0.515
Referencia: ['cahua', 'ilcahua', 'ilcaruaz']
Rank: 3

7 Palabra_ES: teatro
Traducción_NA:
tlamahuiz:0.925
mahuizo:0.723
mahui:0.596
imottilih:0.591
yah:0.569
dral:0.551
diego:0.551
rosario:0.546
itt:0.545
secretario:0.543
Referencia: ['tlamahuiz', 'mahui']
Rank: 1

8 Palabra_ES: alegrar
Traducción_NA:
pahpaqui:0.561
ahuia:0.546
papa:0.525
tehhuellamac:0.508
tlaza:0.503
tinemiz:0.472
cuica:0.458
tixtomahua:0.450
nechico:0.444
nemi:0.444
Referencia: ['pahpaqui', 'ahuia', 'pahpa']
Rank: 1

9 Palabra_ES: ciudad
Traducción_NA:
altepe:0.954
altpe:0.900
hueyi:0.637
huey:0.632
nemiliz:0.590
tlahtoca:0.561
tepe:0.557
mexica:0.552
gali:0.543
huehue:0.507
Referencia: ['altpe', 'altepe']
Rank: 1

10 Palabra_ES: fino
Traducción_NA:
tlazo:0.674
zacuan:0.659
xomoihui:0.650
quetzal:0.620
cahuia:0.613
xical:0.603
quechol:0.597
concui:0.595
pech:0.582
teocuitla:0.579
Referencia: ['tlazo', 'quetzal']
Rank: 1

11 Palabra_ES: leche
Traducción_NA:
ihchican:0.881
notlazohca:0.651
chichi:0.650
itzcuin:0.578
tzitzintin:0.557
illi:0.536
pac:0.528
xmechmo:0.508
chihchi:0.500
xonm:0.498
Referencia: ['chichi']
Rank: 3

12 Palabra_ES: mesero
Traducción_NA:
xexelo:0.704
pech:0.643
ahco:0.643
cualica:0.620
comal:0.501
xeloa:0.497
loyan:0.487
lero:0.476
secretario:0.471
diego:0.471
Referencia: ['xexelo']
Rank: 1

13 Palabra_ES: montaña
Traducción_NA:
huehca:0.701
cem:0.661
tepe:0.626
tlaloa:0.575
miec:0.528
ahquetz:0.522
temo:0.517
```

```
temoa:0.516
miac:0.492
moch:0.491
Referencia: ['tepe']
Rank: 3

14 Palabra_ES: seno
Traducción_NA:
xillan:0.964
notozca:0.841
illam:0.826
itoz:0.805
motozcatlan:0.804
immoma:0.638
uitlapilco:0.609
mocnelili:0.602
potztlatziuh:0.587
timamalihuac:0.574
Referencia: ['xillan']
Rank: 1

15 Palabra_ES: mixcólotl
Traducción_NA:
tepuz:0.751
mixco:0.730
timamalihuac:0.656
tlachia:0.635
tepotz:0.619
cochi:0.618
ahuacatl:0.594
ncmi:0.592
tz:0.584
xuchitl:0.584
Referencia: ['mixco']
Rank: 2

16 Palabra_ES: reunir
Traducción_NA:
potz:0.863
nechico:0.764
puchteca:0.575
nemoa:0.549
pochteca:0.544
xipehua:0.532
iuhqui:0.530
auh:0.528
otitla:0.520
miec:0.517
Referencia: ['cen', 'nechico']
Rank: 2

17 Palabra_ES: madera
Traducción_NA:
icuil:0.700
tlapech:0.641
cuauh:0.558
cuahuitl:0.551
llan:0.538
pech:0.507
ahuacatl:0.489
tepe:0.487
yuhqui:0.483
icpac:0.481
Referencia: ['cuauh', 'tlacuauh', 'cuahuitl']
Rank: 3

18 Palabra_ES: decir
Traducción_NA:
lhuia:0.733
hto:0.729
lhui:0.719
zcuco:0.711
niman:0.711
ilhui:0.708
auh:0.705
quin:0.676
tepahpaquiltia:0.674
zque:0.672
Referencia: ['hto', 'lhuia', 'ilhuia', 'lhui']
Rank: 1

19 Palabra_ES: salvaje
Traducción_NA:
uanime:0.893
ticuahcua:0.624
mexica:0.592
xolochcue:0.591
dral:0.550
secretario:0.541
yb:0.533
rosario:0.532
acamapich:0.527
salveros:0.521
Referencia: ['uanime']
Rank: 1

20 Palabra_ES: echar
Traducción_NA:
patahuaz:0.696
contla:0.688
tz:0.607
zque:0.598
totoca:0.582
quim:0.579
xip:0.577
yb:0.571
```

```
tlaloa:0.567
yaca:0.567
Referencia: ['contla', 'atlatzicuina']
Rank: 2

21 Palabra_ES: pasar
Traducción_NA:
pano:0.918
saroa:0.798
amen:0.631
tlen:0.610
lti:0.581
moch:0.579
omo:0.578
tepahpaquiltia:0.560
yb:0.559
im:0.552
Referencia: ['panoa', 'pano']
Rank: 1

22 Palabra_ES: tomar
Traducción_NA:
zcuco:0.695
xip:0.687
cui:0.681
tepahpaquiltia:0.660
yb:0.648
fa:0.647
dral:0.637
hual:0.631
auh:0.629
zque:0.625
Referencia: ['can', 'cuilia', 'ana', 'cui', 'concui']
Rank: 3

23 Palabra_ES: establecer
Traducción_NA:
itoa:0.580
cauh:0.569
iuh:0.533
acique:0.525
tepe:0.525
pul:0.511
lique:0.503
auh:0.494
chic:0.478
cuezcoma:0.473
Referencia: ['tlalia', 'chan']
Rank: 116

24 Palabra_ES: yo
Traducción_NA:
ne:0.764
nicpia:0.761
fa:0.732
zquia:0.703
hua:0.657
iuh:0.627
e:0.579
tlazohtla:0.575
onie:0.572
mati:0.570
Referencia: ['ne', 'nehhua']
Rank: 1

25 Palabra_ES: familiar
Traducción_NA:
familiar:0.963
xicol:0.639
tlaloa:0.580
chmotlahpalhui:0.580
cocox:0.567
secretario:0.525
salveros:0.523
comal:0.523
dral:0.517
xacualoa:0.517
Referencia: ['familiar']
Rank: 1

26 Palabra_ES: limpiar
Traducción_NA:
chipahuaque:0.905
pohpo:0.811
mitz:0.650
comitl:0.606
stancia:0.584
cciz:0.583
mozacaticonti:0.569
bo:0.563
temi:0.561
caxi:0.559
Referencia: ['chipahuaque', 'popo', 'pohpo']
Rank: 1

27 Palabra_ES: volver
Traducción_NA:
cuepa:0.788
mocuep:0.763
ixpetani:0.703
yahualoa:0.574
ac:0.552
cuep:0.522
cala:0.518
im:0.517
```

```
melauh:0.512
amo:0.510
Referencia: ['cuepa', 'mocuep']
Rank: 1

28 Palabra_ES: dar
Traducción_NA:
dar:0.939
maca:0.846
zcuco:0.691
auh:0.675
mah:0.670
xic:0.665
quin:0.651
taca:0.633
tepahpaquiltia:0.627
omo:0.627
Referencia: ['dar', 'maca']
Rank: 1

29 Palabra_ES: tanto
Traducción_NA:
ih:0.781
zazan:0.683
meh:0.668
amuel:0.627
itemohui:0.616
t:0.589
zcuco:0.574
tepahpaquiltia:0.570
yuhqui:0.555
yb:0.553
Referencia: ['izqui', 'zazan']
Rank: 2

30 Palabra_ES: traer
Traducción_NA:
hual:0.721
cui:0.656
huica:0.645
tih:0.579
tqui:0.570
zcuco:0.561
tepahpaquiltia:0.556
cualica:0.548
t:0.540
xochi:0.536
Referencia: ['huiqui', 'huica', 'cui', 'itqui']
Rank: 2

31 Palabra_ES: escalera
Traducción_NA:
nmediasco:0.899
oquinanquili:0.699
ilpia:0.629
dia:0.609
pualhuany:0.544
comal:0.533
cocone:0.530
mahca:0.522
icuilacta:0.510
nario:0.509
Referencia: ['mama', 'escalera']
Rank: 153

32 Palabra_ES: oír
Traducción_NA:
acica:0.921
nlcaci:0.770
ami:0.663
cac:0.661
cuic:0.636
camati:0.593
tepahpaquiltia:0.562
quimolhui:0.560
mitalhui:0.552
ihuinti:0.548
Referencia: ['cac', 'caqui']
Rank: 4

33 Palabra_ES: comenzar
Traducción_NA:
pehua:0.906
peuh:0.816
pen:0.654
ignacio:0.634
omo:0.577
tlahuiz:0.532
ahuacatl:0.518
yohua:0.514
xip:0.509
dral:0.506
Referencia: ['peuh', 'pehua']
Rank: 1

34 Palabra_ES: itacates
Traducción_NA:
ihtacatl:0.947
huauh:0.660
calhuahza:0.644
pinol:0.597
chichil:0.596
xipehua:0.559
xip:0.546
zacuam:0.534
xoxouhqui:0.530
```

```
castillo:0.527
Referencia: ['ihtacatl']
Rank: 1

35 Palabra_ES: brazo
Traducción_NA:
cuecue:0.650
tonehua:0.571
milacatzo:0.567
icemel:0.566
mama:0.563
timamalihuac:0.548
pahpaqui:0.547
cocone:0.536
ncmi:0.534
huetzi:0.530
Referencia: ['aca']
Rank: 949

36 Palabra_ES: querer
Traducción_NA:
nequi:0.891
znequi:0.778
nequia:0.666
ixquich:0.621
zazo:0.595
fa:0.591
nec:0.589
izta:0.551
den:0.543
amo:0.541
Referencia: ['nequia', 'znequi', 'monequilti', 'nequi']
Rank: 1

37 Palabra_ES: paso
Traducción_NA:
pasilla:0.966
chilcuitlaxcolli:0.745
chil:0.702
zoh:0.654
xumalli:0.603
club:0.597
dral:0.587
secretario:0.581
diego:0.580
federal:0.580
Referencia: ['pasilla']
Rank: 1

38 Palabra_ES: tráiganla
Traducción_NA:
xic:0.826
namoahtapal:0.776
cui:0.688
hual:0.650
yb:0.608
nican:0.596
dar:0.591
necoc:0.574
dral:0.561
diego:0.559
Referencia: ['cui', 'huica']
Rank: 3

39 Palabra_ES: derribar
Traducción_NA:
mota:0.820
tatemp:0.749
tata:0.673
motemilih:0.620
nepapan:0.604
nimitz:0.555
secretario:0.549
dral:0.546
castillo:0.539
tepahpaquiltia:0.538
Referencia: ['tatemp']
Rank: 2

40 Palabra_ES: esfuerzo
Traducción_NA:
quexquich:0.722
chicahua:0.574
ihiyo:0.567
potztlatziuh:0.536
machi:0.510
nohuian:0.495
quezqui:0.493
yb:0.484
tolini:0.482
cahua:0.477
Referencia: ['cuammaca', 'chicahua', 'chicahuaz']
Rank: 2

41 Palabra_ES: unir
Traducción_NA:
f:0.819
george:0.748
estado:0.662
do:0.607
cahuan:0.593
tlahtocayotl:0.582
ahuacatl:0.566
fe:0.554
19:0.552
isoldado:0.533
```

```
Referencia: ['cen']
Rank: 243

42 Palabra_ES: pieza
Traducción_NA:
xexelo:0.818
xeloa:0.612
lero:0.561
loyan:0.542
micuani:0.539
cualica:0.520
nah:0.514
huehca:0.507
tepahpaquiltia:0.495
tlacual:0.495
Referencia: ['xexelo']
Rank: 1

43 Palabra_ES: cultura
Traducción_NA:
cultura:0.972
tlamatiliz:0.961
toyalhuayo:0.719
nahuatlahtol:0.684
lahtol:0.617
iayah:0.610
moyacanilia:0.588
huehue:0.562
moteihtitili:0.556
potz:0.544
Referencia: ['tlamatiliz', 'cultura']
Rank: 1

44 Palabra_ES: cargo
Traducción_NA:
ahqui:0.717
tlaliliz:0.660
chihui:0.660
onino:0.651
im:0.577
xip:0.576
ticmo:0.570
milacatzo:0.568
notelpo:0.565
mix:0.558
Referencia: ['tlahtoca', 'toca']
Rank: 582

45 Palabra_ES: admirar
Traducción_NA:
mahuizo:0.894
imottilih:0.764
tlamahuiz:0.654
itt:0.618
chililiz:0.598
onimitz:0.583
mahui:0.571
mihtlanililia:0.561
itti:0.541
xip:0.538
Referencia: ['tlamahuiz', 'mahuizo']
Rank: 1

46 Palabra_ES: pez
Traducción_NA:
mimic:0.834
michi:0.642
ahhua:0.627
ictique:0.615
dral:0.600
diego:0.593
mic:0.592
secretario:0.584
castillo:0.580
rosario:0.576
Referencia: ['michi','mimic']
Rank: 1

47 Palabra_ES: valer
Traducción_NA:
1:0.675
quexquich:0.650
ton:0.596
mama:0.494
yancui:0.484
quezqui:0.475
toma:0.471
s:0.469
nuevo:0.462
ki:0.461
Referencia: ['patio']
Rank: 14

48 Palabra_ES: caja
Traducción_NA:
petlacal:0.932
top:0.752
tlachie:0.591
ixtelolo:0.524
ilpi:0.522
ui:0.510
cxi:0.501
xillan:0.500
pie:0.492
mohuaxca:0.491
Referencia: ['petlacal']
```

```
Rank: 1

49 Palabra_ES: vender
Traducción_NA:
tlanamacazquia:0.957
tlanamaca:0.956
namaco:0.818
namaca:0.792
villistas:0.626
rosario:0.624
secretario:0.617
dral:0.610
salveros:0.606
diego:0.606
Referencia: ['tlanamacazquia', 'namaca', 'tlanamaca', 'namaco']
Rank: 1

50 Palabra_ES: rostro
Traducción_NA:
mixco:0.816
tlachia:0.752
yollo:0.691
timamalihuac:0.685
notelpo:0.657
mix:0.644
notozca:0.634
ixeo:0.634
naz:0.622
chihui:0.620
Referencia: ['mixco'. 'ix', 'mix']
Rank: 1

51 Palabra_ES: descanso
Traducción_NA:
mocehui:0.848
calyolot:0.636
huizque:0.582
pepe:0.579
immotla:0.570
icpal:0.561
acia:0.554
tlauh:0.517
timamalihuac:0.516
ciauh:0.511
Referencia: ['mocehui']
Rank: 1

52 Palabra_ES: lodo
Traducción_NA:
atlatzicuina:0.910
zoqui:0.689
atahuit:0.643
mimacaci:0.640
itti:0.625
caxi:0.593
ich:0.554
tema:0.543
salvador:0.531
motzicoa:0.524
Referencia: ['zoqui']
Rank: 2

53 Palabra_ES: perdonar
Traducción_NA:
pohpo:0.909
mozacaticonti:0.759
otimitzno:0.702
chipahuaque:0.665
pahcayo:0.599
mochantilia:0.515
comitl:0.513
stancia:0.501
iahhuui:0.488
itzoltilia:0.482
Referencia: ['pohpo', 'popol']
Rank: 1

54 Palabra_ES: pipián
Traducción_NA:
chilacach:0.965
temolcax:0.654
textli:0.629
secretario:0.626
castillo:0.619
acamapich:0.616
yohuac:0.614
rosario:0.614
rafael:0.611
salveros:0.610
Referencia: ['yohuac', 'chilacach']
Rank: 1

55 Palabra_ES: llanura
Traducción_NA:
xtlahua:0.820
as:0.623
ip:0.586
ua:0.567
mitz:0.537
tlahui:0.528
xinach:0.507
tlalli:0.504
eu:0.449
zaca:0.448
Referencia: ['xtlahua','tlal', 'tlalli']
Rank: 1
```

```
56 Palabra_ES: transportar
Traducción_NA:
axiti:0.750
chalchiuh:0.688
chalchihui:0.660
quetzal:0.626
illam:0.604
zacuan:0.565
itoz:0.561
xillan:0.553
teu:0.548
xip:0.547
Referencia: ['axiti']
Rank: 1

57 Palabra_ES: ver
Traducción_NA:
ver:0.945
tta:0.812
yb:0.647
uh:0.641
xip:0.634
auh:0.626
tac:0.618
cen:0.614
zcuco:0.612
oncan:0.607
Referencia: ['tta', 'itac', 'ver', 'conmottilia', 'itta']
Rank: 1

58 Palabra_ES: vasija
Traducción_NA:
comitl:0.727
cax:0.694
chiquihui:0.655
tecoma:0.638
zoqui:0.627
caxi:0.612
nechicalhui:0.612
xommo:0.566
xaxa:0.559
cuitlahui:0.540
Referencia: ['comitl', 'cax', 'tecoma']
Rank: 1

59 Palabra_ES: sabor
Traducción_NA:
mati:0.622
lan:0.598
tro:0.562
chil:0.549
iahhuui:0.500
ermanos:0.478
yol:0.477
choh:0.471
mian:0.467
ahnozo:0.465
Referencia: ['iahhuui', 'mati']
Rank: 1

60 Palabra_ES: amado
Traducción_NA:
tlazohtla:0.935
ama:0.794
ticcua:0.793
zquia:0.758
tlazo:0.643
diego:0.588
tazohta:0.582
comal:0.581
niz:0.575
federal:0.571
Referencia: ['tlazo', 'tlazohtla']
Rank: 1

61 Palabra_ES: cofre
Traducción_NA:
top:0.939
petlacal:0.844
xillan:0.557
itoz:0.557
ilpi:0.553
tlachie:0.544
motozcatlan:0.524
pale:0.523
xuchitl:0.521
onimitz:0.518
Referencia: ['top', 'petlacal']
Rank: 1

62 Palabra_ES: tigre
Traducción_NA:
ocelo:0.956
tecuan:0.721
zacuan:0.705
quechol:0.656
cacehuaztli:0.628
totome:0.568
quetzal:0.568
ehua:0.566
dral:0.546
tototl:0.541
Referencia: ['ocelo', 'tecuan']
Rank: 1

63 Palabra_ES: ramo
Traducción_NA:
```

```
ramos:0.927
g:0.704
señor:0.595
pixca:0.546
tlaca:0.542
gr:0.536
ri:0.530
vor:0.523
poalli:0.523
egorio:0.495
Referencia: ['ramos']
Rank: 1

64 Palabra_ES: comer
Traducción_NA:
cua:0.815
cuaz:0.814
tos:0.731
tlacual:0.721
cualtiz:0.661
tepahpaquiltia:0.607
zque:0.586
dral:0.550
cuap:0.549
niquintlayocoli:0.546
Referencia: ['cuaz', 'cualtiz', 'otlacua', 'cua', 'tlacual']
Rank: 1

65 Palabra_ES: cierto
Traducción_NA:
melahua:0.943
nel:0.747
mitzmo:0.578
hualla:0.574
yb:0.572
ichpoc:0.569
xip:0.563
milacatzo:0.563
ihtoh:0.562
zque:0.560
Referencia: ['melahua', 'nel', 'cecni']
Rank: 1

66 Palabra_ES: tierno
Traducción_NA:
xoxouhqui:0.772
stancia:0.617
calhuahza:0.599
chilmolli:0.598
nohpal:0.595
xoxo:0.583
nopal:0.573
catarina:0.550
izhua:0.543
ahnozo:0.540
Referencia: ['xoxouhqui', 'ne']
Rank: 1

67 Palabra_ES: lavar
Traducción_NA:
paca:0.928
caxi:0.747
tlalia:0.669
tlahuicxitilli:0.663
ihti:0.640
motlali:0.630
mochi:0.627
huic:0.624
tec:0.619
mp:0.618
Referencia: ['paca']
Rank: 1

68 Palabra_ES: gritar
Traducción_NA:
tzatzi:0.959
lero:0.560
ihcahuaca:0.552
tzahtzi:0.531
dral:0.464
secretario:0.463
diego:0.463
tlacacalata:0.462
salveros:0.460
acamapich:0.460
Referencia: ['tzatzi', 'tzahtzi']
Rank: 1

69 Palabra_ES: jugar
Traducción_NA:
mahui:0.901
nocniu:0.682
ltih:0.668
ahuil:0.607
mahuizzo:0.575
pahpa:0.561
ihcuilo:0.528
cuetlaxo:0.522
sombra:0.516
tizque:0.513
Referencia: ['mahui']
Rank: 1

70 Palabra_ES: enorme
Traducción_NA:
```

```
mil:0.766
huehca:0.717
cpac:0.619
tepahpaquiltia:0.598
jard:0.596
xochi:0.594
dral:0.584
zcuco:0.584
occe:0.583
moch:0.580
Referencia: ['huey']
Rank: 277

71 Palabra_ES: anáhuac
Traducción_NA:
noch:0.713
nahuac:0.660
tzaoctimanca:0.657
siempre:0.649
tlamatiliz:0.605
cultura:0.603
m:0.579
dueño:0.578
zcuco:0.567
de:0.563
Referencia: ['nahuac']
Rank: 2

72 Palabra_ES: acabar
Traducción_NA:
tlami:0.709
mic:0.661
cauh:0.646
ahui:0.623
yoltequipacho:0.595
auh:0.590
miqui:0.566
tami:0.562
totecuyo:0.550
tepahpaquiltia:0.537
Referencia: ['tami', 'tlami']
Rank: 1

73 Palabra_ES: pluma
Traducción_NA:
quetzal:0.915
ihui:0.799
potonia:0.725
tapayol:0.674
huitz:0.672
xahua:0.670
zacuan:0.654
coatl:0.649
ten:0.643
quechol:0.630
Referencia: ['quetzal']
Rank: 1

74 Palabra_ES: indígena
Traducción_NA:
lahtol:0.874
macehual:0.796
nahuatlahtol:0.643
toyalhuayo:0.574
moyacanilia:0.567
elehuia:0.562
cultura:0.539
tlamatiliz:0.518
macihui:0.515
español:0.512
Referencia: ['macehual']
Rank: 2

75 Palabra_ES: adornar
Traducción_NA:
uan:0.651
chihchihua:0.642
apana:0.607
quimon:0.544
oquich:0.536
yahualoa:0.525
xip:0.502
tizque:0.497
tlauhtia:0.497
oztomeca:0.490
Referencia: ['chihchihua', 'apana']
Rank: 2

76 Palabra_ES: presidente
Traducción_NA:
presi:0.783
us:0.732
tlahtoca:0.693
ludovi:0.651
premios:0.600
tlahtoani:0.591
hueyi:0.589
nario:0.579
yacana:0.566
gali:0.565
Referencia: ['presi']
Rank: 1

77 Palabra_ES: quiltoniles
```

```
Traducción_NA:
cualac:0.895
illi:0.812
alac:0.690
cuacualaca:0.604
apiaz:0.581
1:0.575
oomicihcicu:0.575
chilmolli:0.551
quil:0.549
chil:0.545
Referencia: ['quil']
Rank: 9

78 Palabra_ES: aire
Traducción_NA:
yehyecatl:0.764
eheca:0.693
eca:0.656
tetzahui:0.640
hui:0.623
mach:0.571
tlacochcalca:0.553
yeyeca:0.549
huilo:0.540
xip:0.534
Referencia: ['yehyecatl', 'yeyeca', 'eheca']
Rank: 1

79 Palabra_ES: barca
Traducción_NA:
acalli:0.829
latuic:0.582
ltemalo:0.575
ono:0.540
tlatia:0.536
tzitz:0.498
choqui:0.471
poyahua:0.469
aquia:0.466
tlathui:0.457
Referencia: ['acalli']
Rank: 1

80 Palabra_ES: conocer
Traducción_NA:
ixmati:0.842
quix:0.762
mati:0.738
iximacho:0.620
machi:0.601
ixmachilia:0.597
quiximati:0.593
choh:0.568
tlaca:0.567
y:0.560
Referencia: ['ixmachilia', 'iximacho', 'oixmatilia', 'ixmati', 'mati']
Rank: 1

81 Palabra_ES: muerto
Traducción_NA:
mimic:0.624
mic:0.613
ictique:0.560
cuecuepotza:0.552
im:0.548
miqui:0.535
oquin:0.533
potoni:0.530
tzontequili:0.524
yb:0.515
Referencia: ['mimic', 'mic']
Rank: 1

82 Palabra_ES: silvestre
Traducción_NA:
tlal:0.808
tlalticpac:0.665
ayo:0.593
tal:0.572
y:0.563
ax:0.553
talmanic:0.541
al:0.539
tin:0.526
cuechac:0.525
Referencia: ['cuauh', 'tlal']
Rank: 1

83 Palabra_ES: final
Traducción_NA:
mix:0.799
xip:0.709
im:0.691
yb:0.682
dral:0.674
mochi:0.670
diego:0.655
cauh:0.653
zcuco:0.643
y:0.642
Referencia: ['tlatz']
Rank: 57

84 Palabra_ES: suceder
```

```
Traducción_NA:
chiuh:0.563
chihua:0.516
cholo:0.513
xip:0.502
ga:0.488
oquimo:0.482
quenin:0.476
tlaca:0.476
cequin:0.475
tihuetz:0.473
Referencia: ['chiuh', 'chiu', 'pano', 'panoa']
Rank: 1

85 Palabra_ES: venir
Traducción_NA:
hual:0.883
hualla:0.717
huitz:0.659
m:0.621
yb:0.616
zcuco:0.615
zque:0.597
quin:0.578
auh:0.574
dral:0.571
Referencia: ['hual', 'hualla']
Rank: 1

86 Palabra_ES: molcajete
Traducción_NA:
molcaxitl:0.971
temolcax:0.863
otzoy:0.687
caxi:0.684
xommo:0.679
zoqui:0.660
tlaxamanil:0.633
tzoaloni:0.622
chiquihui:0.590
stancia:0.589
Referencia: ['molcaxitl', 'temolcax']
Rank: 1

87 Palabra_ES: pantalón
Traducción_NA:
pantalon:0.980
beethoven:0.771
chachal:0.767
villistas:0.761
comal:0.757
echicoliz:0.756
anechhuall:0.752
matiloa:0.751
rafael:0.746
acamapich:0.739
Referencia: ['pantalon']
Rank: 1

88 Palabra_ES: aguacate
Traducción_NA:
aztlacapalli:0.948
tolachno:0.814
ahua:0.668
pix:0.593
poyox:0.570
diego:0.505
yb:0.500
dral:0.497
izhua:0.496
rosario:0.489
Referencia: ['ahua']
Rank: 3

89 Palabra_ES: tepaneca
Traducción_NA:
colhuaca:0.886
eca:0.682
popolo:0.627
culhuacan:0.601
oquim:0.578
ahuacatl:0.550
mexica:0.546
motlanahuatili:0.540
cocox:0.528
quim:0.523
Referencia: ['eca', 'tepan']
Rank: 2

90 Palabra_ES: coyote
Traducción_NA:
camachalo:0.670
coyoc:0.612
ichca:0.601
cuaz:0.537
zauh:0.524
huitequi:0.522
zahua:0.504
pu:0.495
extiz:0.488
dios:0.471
Referencia: ['coyo', 'coyoc']
Rank: 2
```

```
91 Palabra_ES: mirada
Traducción_NA:
tlachia:0.800
im:0.724
mix:0.701
xip:0.689
tz:0.677
yuhqui:0.653
milacatzo:0.648
ix:0.647
tta:0.644
auh:0.641
Referencia: ['tlachia']
Rank: 1

92 Palabra_ES: chiquillo
Traducción_NA:
ninozcalia:0.781
cuiya:0.609
xip:0.604
cuitl:0.586
miec:0.579
ten:0.579
mahui:0.578
pohua:0.570
itzopelica:0.565
immoma:0.561
Referencia: ['pipil']
Rank: 454

93 Palabra_ES: salvador
Traducción_NA:
salvador:0.949
tema:0.808
quixti:0.731
pr:0.564
cachihualiz:0.556
jesu:0.546
macehui:0.525
xip:0.511
yia:0.509
yb:0.508
Referencia: ['quixti', 'salvador', 'tema']
Rank: 1

94 Palabra_ES: puramente
Traducción_NA:
ommihzo:0.807
imayauhcan:0.778
imayauhcampa:0.638
ceuh:0.574
nepantla:0.561
iye:0.544
dral:0.536
pochcopa:0.535
secretario:0.532
tepahpaquiltia:0.528
Referencia: ['zan', 'za']
Rank: 69

95 Palabra_ES: llevar
Traducción_NA:
huica:0.870
mama:0.686
cuica:0.651
tqui:0.632
polactia:0.600
zcuco:0.599
tepahpaquiltia:0.596
xip:0.596
cui:0.592
quiliz:0.587
Referencia: ['cuica', 'tqui', 'huica']
Rank: 1

96 Palabra_ES: ropa
Traducción_NA:
tzotzoma:0.979
ui:0.605
dral:0.547
nuevo:0.541
tzatzapal:0.539
xolochcue:0.537
diego:0.534
secretario:0.534
rosario:0.533
acamapich:0.532
Referencia: ['tzotzoma']
Rank: 1

97 Palabra_ES: tema
Traducción_NA:
tema:0.956
salvador:0.769
zcal:0.638
quixti:0.604
cachihualiz:0.602
xip:0.600
auh:0.597
zque:0.591
zcuco:0.585
yuhqui:0.576
Referencia: ['tema']
```

```
Rank: 1

98 Palabra_ES: rogar
Traducción_NA:
tahtan:0.909
tlauhti:0.891
tamechtahtanil:0.777
tlauhtia:0.679
ili:0.647
idl:0.622
xip:0.582
nimitz:0.565
end:0.562
namech:0.559
Referencia: ['tahtan']
Rank: 1

99 Palabra_ES: vestido
Traducción_NA:
tzotzoma:0.660
nequia:0.598
sa:0.569
xolochcue:0.544
quen:0.539
cocone:0.518
cah:0.505
ilpi:0.477
icuilacta:0.472
oncuan:0.467
Referencia: ['tzotzoma', 'quen']
Rank: 1

100 Palabra_ES: indio
Traducción_NA:
ig:0.738
macehual:0.703
quixtiz:0.607
quiliz:0.602
obispo:0.560
xip:0.547
monexiti:0.541
iteopixca:0.540
ocaxil:0.540
tta:0.538
Referencia: ['macehual']
Rank: 2

101 Palabra_ES: llano
Traducción_NA:
talmanic:0.927
mani:0.699
tlal:0.638
tlalticpac:0.628
dral:0.561
diego:0.551
secretario:0.546
rosario:0.545
castillo:0.537
acamapich:0.537
Referencia: ['xtlahua', 'tlal', 'talmanic']
Rank: 1

102 Palabra_ES: descubrir
Traducción_NA:
ixpetla:0.828
nexti:0.655
tolini:0.646
timamalihuac:0.583
mix:0.576
cala:0.550
namiqui:0.543
tlachia:0.519
tz:0.517
ixeo:0.512
Referencia: ['ixpetla', 'nexti']
Rank: 1

103 Palabra_ES: niño
Traducción_NA:
tzitzintin:0.786
cocone:0.772
conetl:0.739
pipil:0.731
toton:0.702
pil:0.682
cihuanton:0.680
machti:0.600
amen:0.587
telpoca:0.583
Referencia: ['pipil', 'cihuanton', 'cocone', 'conetl']
Rank: 2

104 Palabra_ES: desierto
Traducción_NA:
ceuh:0.801
er:0.781
des:0.770
ommihzo:0.606
olpi:0.569
izquin:0.549
mp:0.478
diego:0.475
dral:0.468
achtopa:0.463
Referencia: ['des']
```

```
Rank: 3

105 Palabra_ES: efecto
Traducción_NA:
ihtotia:0.625
tlaocox:0.613
oquich:0.593
nehnemi:0.564
mixco:0.518
ellel:0.499
tz:0.499
mito:0.495
oceloquichtle:0.476
quimon:0.473
Referencia: ['nel']
Rank: 777

106 Palabra_ES: diverso
Traducción_NA:
nepapan:0.874
mota:0.832
motemilih:0.764
tatemp:0.561
huipan:0.557
tepahpaquiltia:0.531
huiptla:0.521
dral:0.506
diego:0.502
secretario:0.500
Referencia: ['nepapan']
Rank: 1

107 Palabra_ES: coger
Traducción_NA:
ana:0.638
ochol:0.567
cholo:0.562
milacatzo:0.560
tecuan:0.549
cuecue:0.540
ictique:0.537
tzicuin:0.524
mac:0.516
choloa:0.514
Referencia: ['ana', 'cui', 'cana']
Rank: 1

108 Palabra_ES: vaca
Traducción_NA:
uacax:0.940
enepil:0.819
huen:0.657
cuacua:0.621
cueh:0.617
tictohuaxcati:0.541
cihua:0.520
mp:0.503
panoa:0.479
dral:0.477
Referencia: ['uacax', 'cuacua']
Rank: 1

109 Palabra_ES: quitar
Traducción_NA:
ihcuani:0.687
cciz:0.681
nech:0.648
cuilia:0.611
cuih:0.582
tlalia:0.562
cui:0.550
huizque:0.542
pachoa:0.532
mochi:0.530
Referencia: ['cuicuil', 'cuilia']
Rank: 4

110 Palabra_ES: presencia
Traducción_NA:
tlacuauh:0.632
mopechteccac:0.631
tzinco:0.598
quimolhui:0.597
tlazoh:0.550
contla:0.549
tzicuin:0.542
eltilitiuh:0.542
uad:0.537
ix:0.524
Referencia: ['ix']
Rank: 10

111 Palabra_ES: pídele
Traducción_NA:
htlan:0.915
puebla:0.764
mayana:0.599
illi:0.578
htlani:0.568
icni:0.551
apiz:0.550
xicol:0.547
apiaz:0.533
secretario:0.521
Referencia: ['htlan', htlani]
Rank: 1
```

```
112 Palabra_ES: pedirle
Traducción_NA:
mihtlanililia:0.955
pouhilil:0.759
neltili:0.669
tti:0.659
titlani:0.650
chiyali:0.644
chililiz:0.644
eltilitiuh:0.643
quimolhui:0.628
imottilih:0.625
Referencia: ['mihtlanililia', 'htlan', htlani]
Rank: 1

113 Palabra_ES: izquierda
Traducción_NA:
pochcopa:0.925
poch:0.624
imayauhcampa:0.592
imayauhcan:0.515
au:0.505
mian:0.502
acacic:0.501
iye:0.494
diego:0.479
dral:0.479
Referencia: ['pochcopa', 'opoch', 'poch']
Rank: 1

114 Palabra_ES: ocurrir
Traducción_NA:
yoh:0.641
quem:0.636
pano:0.581
machi:0.563
zquiani:0.546
onino:0.538
ammo:0.530
bar:0.529
nicnemili:0.529
chiuh:0.522
Referencia: ['chiuh', 'pano']
Rank: 3

115 Palabra_ES: despreciar
Traducción_NA:
ech:0.785
xinech:0.770
tlecuezal:0.658
pu:0.600
cizu:0.592
ilama:0.562
pehua:0.558
xmechmo:0.548
tlachie:0.548
dral:0.538
Referencia: ['mah', 'pehua']
Rank: 7

116 Palabra_ES: agregar
Traducción_NA:
motlali:0.846
tiliaya:0.821
tlahuicxitilli:0.713
tlalia:0.678
cuacualaca:0.673
tec:0.649
itzoltilia:0.641
huic:0.632
paca:0.597
caxi:0.596
Referencia: ['motlali', 'tlalia']
Rank: 1

117 Palabra_ES: amargo
Traducción_NA:
chichi:0.662
itzcuin:0.606
choca:0.529
miz:0.521
tepahpaquiltia:0.521
onohualco:0.513
mimizt:0.509
ixpetla:0.503
pia:0.501
xip:0.499
Referencia: ['chichi']
Rank: 1

118 Palabra_ES: centenario
Traducción_NA:
nario:0.914
premios:0.821
icuilacta:0.795
ipanoque:0.739
presi:0.695
tazohta:0.649
hcuilo:0.601
elehui:0.597
diego:0.590
dral:0.586
Referencia: ['nario']
```

```
Rank: 1

119 Palabra_ES: voz
Traducción_NA:
tzahtzi:0.885
nopiltzin:0.595
xinechmotlapohpolhuili:0.539
mu:0.529
omitz:0.521
tzatzi:0.505
im:0.498
tepotz:0.494
dral:0.492
tze:0.487
Referencia: ['tzatzi', 'tzahtzi']
Rank: 1

120 Palabra_ES: animal
Traducción_NA:
yolca:0.972
v:0.703
pu:0.589
iapizmi:0.555
pah:0.548
mictia:0.534
rn:0.512
ichca:0.509
ea:0.502
miltequitca:0.500
Referencia: ['yoyolcameh', 'yolca']
Rank: 1

121 Palabra_ES: nombre
Traducción_NA:
toca:0.867
tocayoti:0.716
tlahtoca:0.629
hispan:0.626
v:0.612
de:0.611
ixiptla:0.600
y:0.599
gali:0.593
año:0.591
Referencia: ['toca']
Rank: 1

122 Palabra_ES: pedir
Traducción_NA:
htlani:0.849
tlania:0.633
tahtan:0.616
tamechtahtanil:0.610
neltili:0.605
nimitz:0.588
lih:0.568
ichpoc:0.537
htlan:0.527
tenehui:0.512
Referencia: ['htlani', 'htlan', 'mihtlanililia', 'tahtan', 'tamechtahtanil']
Rank: 1

123 Palabra_ES: nabo
Traducción_NA:
stancia:0.922
bo:0.894
catarina:0.792
ga:0.675
cuechahua:0.648
cec:0.586
pexon:0.566
temolcax:0.556
zul:0.553
chipahuaque:0.551
Referencia: ['bo']
Rank: 2

124 Palabra_ES: derecho
Traducción_NA:
ihtoh:0.816
melahua:0.727
iyaca:0.677
melauh:0.664
ozto:0.624
nel:0.582
milacatzo:0.531
itt:0.524
tecu:0.509
pilca:0.509
Referencia: ['melahua', 'yec', 'melauh']
Rank: 2

125 Palabra_ES: uña
Traducción_NA:
cuicuil:0.799
moci:0.721
ahuiz:0.660
poch:0.581
mihiyohuilti:0.545
itz:0.541
diego:0.518
tlaloa:0.515
dral:0.507
ciauh:0.504
Referencia: ['itz']
Rank: 6
```

```
126 Palabra_ES: naranja
Traducción_NA:
naranja:0.975
juzga:0.615
ameyal:0.571
tzopel:0.556
matiloa:0.535
comal:0.533
orcasitas:0.532
ieltapach:0.531
chachal:0.528
perah:0.522
Referencia: ['xal', 'naranja']
Rank: 1

127 Palabra_ES: prestar
Traducción_NA:
ehuia:0.937
onite:0.754
icni:0.584
jo:0.538
19:0.531
nocupix:0.531
diego:0.523
rosario:0.514
dral:0.512
acamapich:0.506
Referencia: ['ehuia']
Rank: 1

128 Palabra_ES: perforar
Traducción_NA:
coyoni:0.941
coyoc:0.790
acacic:0.690
ohuehyiyac:0.626
poch:0.569
amaihcuilol:0.559
rdenas:0.553
diego:0.542
dral:0.539
tecu:0.535
Referencia: ['coyoni', 'coyoc']
Rank: 1

129 Palabra_ES: portal
Traducción_NA:
ticchi:0.710
por:0.596
chia:0.585
min:0.570
yehhua:0.556
tlalticpac:0.521
axcan:0.519
moyacanilia:0.514
chien:0.503
ohuala:0.501
Referencia: ['por', 'tal']
Rank: 2

130 Palabra_ES: introducir
Traducción_NA:
calaqui:0.747
calaquia:0.718
calac:0.686
nican:0.610
xip:0.562
hual:0.543
puerta:0.511
im:0.502
cui:0.497
auh:0.492
Referencia: ['calaqui', 'calaquia']
Rank: 1

131 Palabra_ES: quetzal
Traducción_NA:
quetzal:0.982
coatl:0.729
zacuan:0.722
chalchihui:0.711
chalchiuh:0.682
ihui:0.663
potonia:0.656
quechol:0.637
tapayol:0.634
cozcatl:0.634
Referencia: ['quetzal']
Rank: 1

132 Palabra_ES: alimento
Traducción_NA:
pal:0.690
tota:0.677
nuestro:0.618
ol:0.604
celi:0.604
tlacual:0.581
quiyauh:0.575
ihchican:0.561
xip:0.551
```

```
toh:0.547
Referencia: ['atetzocoa', 'pal']
Rank: 1

133 Palabra_ES: tapar
Traducción_NA:
pachoa:0.953
cciz:0.738
cuih:0.727
motequipacho:0.693
pacho:0.662
itzoltilia:0.653
zoqui:0.647
tlahuicxitilli:0.614
caxi:0.614
tequi:0.609
Referencia: ['pachoa']
Rank: 1

134 Palabra_ES: tortuga
Traducción_NA:
tectli:0.816
tlahuicxitilli:0.678
ayo:0.676
caxi:0.633
tlalia:0.628
zoqui:0.626
paca:0.623
alac:0.610
xip:0.607
chiyahuizotl:0.592
Referencia: ['ayo']
Rank: 3

135 Palabra_ES: levantar
Traducción_NA:
quetza:0.634
yb:0.597
ohuatzintic:0.592
yollo:0.560
xip:0.540
bar:0.539
zcuco:0.537
quechi:0.530
dral:0.527
diego:0.524
Referencia: ['quetza']
Rank: 1

136 Palabra_ES: verde
Traducción_NA:
xoxo:0.883
xoxouhqui:0.839
ayohhuach:0.715
zaca:0.708
chilmolli:0.657
molli:0.623
calhuahza:0.568
chalchiuh:0.563
xole:0.556
chichil:0.547
Referencia: ['xoxouhqui', 'xoxo']
Rank: 1

137 Palabra_ES: grande
Traducción_NA:
hueyi:0.955
huey:0.693
hue:0.656
altpe:0.623
gali:0.601
ohuehyiyac:0.597
altepe:0.581
chamahuac:0.576
izcal:0.567
piloa:0.559
Referencia: ['hueyi', 'huey']
Rank: 1

138 Palabra_ES: alfredo
Traducción_NA:
alfredo:0.976
matiloa:0.767
chachal:0.760
felipe:0.735
xacualoa:0.730
ieltapach:0.730
salveros:0.727
club:0.727
rafael:0.726
organil:0.725
Referencia: ['alfredo']
Rank: 1

139 Palabra_ES: sólo
Traducción_NA:
zan:0.851
zcuco:0.749
im:0.742
choh:0.730
xip:0.696
yuhqui:0.693
```

```
yb:0.679
zque:0.676
auh:0.669
omo:0.664
Referencia: ['zan', 'za']
Rank: 1

140 Palabra_ES: patio
Traducción_NA:
patio:0.978
itual:0.969
ithual:0.922
nepantla:0.650
ber:0.634
calitic:0.564
nex:0.514
pixca:0.511
ticatca:0.507
cotona:0.503
Referencia: ['itual', 'patio', 'ithual']
Rank: 1

P@1= 0.671
P@5= 0.886
P@10 0.914
```

Apéndice F

Listado de publicaciones

- **Memorias de congresos y revistas**

Ximena Gutierrez-Vasques. 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*. pages 154–160.

Ximena Gutiérrez-Vasques, Elena Carolina Vilchis Vargas, and Cerbón Ynclán Rocío. 2015. Recopilación de un corpus paralelo electrónico para una lengua minoritoria: el caso del nahuatl-español. In *Primer Congreso Internacional el Patrimonio Cultural y las Nuevas Tecnologías*. INAH.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Ximena Gutierrez-Vasques. 2017. Exploring bilingual lexicon extraction for Spanish-Nahuatl. In *ACL Workshop in Women and Underrepresenting Minorities in Natural Language Processing*.

Ximena Gutierrez-Vasques and Victor Mijangos. 2017. Low-resource bilingual lexicon extraction using graph based word embeddings. *arXiv preprint arXiv:1710.02569 (Accepted in International Journal of Computational Linguistics and Application)*.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. In *The 27th International Conference on Computational Linguistics (COLING 2018)*.

- **Capítulos de libro**

Ximena Gutierrez Vasques (2018). Corpus paralelo español-náhuatl y su uso en las tecnologías del lenguaje humano In Galina Russell, Isabel; Peña Pimentel, Miriam; Priani Saisó, Ernesto; Barrón Tovar, José Francisco; Domínguez Herbón, David; Álvarez Sánchez, Adriana (Coords), *Humanidades digitales: lengua, texto, patrimonio y datos*. Mexico, Bonilla Artigas Editores.

Bibliografía

Oliver Adams, Graham Neubig, Trevor Cohn, y Steven Bird. 2015. Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. In *12th International Workshop on Spoken Language Translation (IWSLT)*.

Lars Ahrenberg, Mikael Andersson, y Magnus Merkel. 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 29–35.

Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein, y Jörg Tiedemann. 2000. Evaluation of word alignment systems. In *LREC*. volume 2000, pages 1255–1261.

Ethem Alpaydin. 2016. *Machine Learning: The New AI*. MIT Press.

Vivien Altmann y Gabriel Altmann. 2008. Anleitung zu quantitativen textanalysen. *Methoden und Anwendungen* .

Maryam Aminian, Mohammad Sadegh Rasooli, y Mona Diab. 2017. Transferring semantic roles using translation and syntactic information. *arXiv preprint arXiv:1710.01411* .

Stephen R Anderson. 2015. Dimensions of morphological complexity. *Understanding and measuring morphological complexity* pages 11–26.

Mikel Artetxe, Gorka Labaka, y Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2289–2294.

Mikel Artetxe, Gorka Labaka, y Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Matthew Baerman, Dunstan Brown, y Greville G Corbett. 2015. *Understanding and measuring morphological complexity*. Oxford University Press, USA.

Marco Baroni, Raffaela Bernardi, y Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)* 9.

Marco Baroni y Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4):673–721.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, y Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

Christian Bentz, Tatjana Soldatova, Alexander Koplenig, y Tanja Samardžić. 2016. A comparison between morphological complexity measures: typological data vs. language corpora .

Phil Blunsom y Karl Moritz Hermann. 2014. Multilingual distributed representations without word alignment .

Piotr Bojanowski, Edouard Grave, Armand Joulin, y Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Ignacio Bosque. 1983. La morfología. *Introducción a la Lingüística. Madrid* .

Jan Botha y Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning*. pages 1899–1907.

Martin Braschler y Peter Scäuble. 1998. Multilingual information retrieval based on document alignment techniques. In *Research and Advanced Technology for Digital Libraries*, Springer, pages 183–197.

Peter F Brown, Jennifer C Lai, y Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 169–176.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, y Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19(2):263–311.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, y Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203* .

Mehmet Talha Cakmak, Süleyman Acar, y Gülsen Eryigit. 2012. Word alignment for english-turkish language pair. In *LREC*. pages 2177–2180.

Horacio Carochi. 1645. Arte de la lengua mexicana con la declaración de todos sus adverbios.

Kenneth Ward Church y Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1):22–29.

Ryan Cotterell, Sebastian J Mielke, Jason Eisner, y Brian Roark. 2018. Are all languages equally hard to language-model? *arXiv preprint arXiv:1806.03743* .

Mathias Creutz y Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, y Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6):391.

Hervé Déjean, Éric Gaussier, y Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 1–7.

Mona Diab y Steve Finch. 2000. A statistical word-level translation model for comparable corpora. Technical report, DTIC Document.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* 10(7):1895–1923. https://doi.org/10.1162/089976698300017197.

Georgiana Dinu, Angeliki Lazaridou, y Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568* .

Wolfgang U Dressler. 1985. Suppletion in word formation. *Historical semantics—historical word-formation* pages 97–112.

A El-Desoky Mousa, H-KJ Kuo, Lidia Mangu, y Hagen Soltau. 2013. Morpheme-based feature-rich language models using deep neural networks for lvcsr of egyptian arabic. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, pages 8435–8439.

Antonio Fábregas. 2013. *La morfología: el análisis de la palabra compleja*. Ed. Síntesis.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* .

Pascale Fung. 2000. A statistical view on bilingual lexicon extraction. In *Parallel Text Processing*, Springer, pages 219–236.

William A Gale y Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics* 19(1):75–102.

Dan Garrette y Hannah Alpert-Abrams. 2016. An unsupervised model of orthographic variation for historical document transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 467–472.

Eric Gaussier, J-M Renders, Irina Matveeva, Cyril Goutte, y Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, page 526.

Alexander Gelbukh y Grigori Sidorov. 2006. Alignment of paragraphs in bilingual texts using bilingual dictionaries and dynamic programming. In *Progress in Pattern Recognition, Image Analysis and Applications*, Springer, pages 824–833.

Alexander Gelbukh, Grigori Sidorov, y Grigori Sidorov Alexander Gelbukh. 2006. Procesamiento automático del español con enfoque en recursos léxicos grandes. Technical report, e-libro, Corp.

Alexander Gelbukh, Grigori Sidorov, Diego Lara-Reyes, y Liliana Chanona-Hernandez. 2008. Division of spanish words into morphemes with a genetic algorithm. In *International Conference on Application of Natural Language to Information Systems*. Springer, pages 19–26.

Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, y Kenji Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 228–235.

Xavier Gómez Guinovart. 2012. A hybrid corpus-based approach to bilingual terminology extraction. *Encoding the Past, Decoding The Future: Corpora in the 21st Century* pages 147–175.

Thomas L Griffiths, Mark Steyvers, y Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review* 114(2):211.

Aditya Grover y Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 855–864.

Peter Grunwald. 2004. A tutorial introduction to the minimum description length principle. *arXiv preprint math/0406077* .

Ximena Gutierrez-Vasques, Gerardo Sierra, y Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, y Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*. volume 2008, pages 771–779.

Zellig S Harris. 1954. Distributional structure. *Word* 10(2-3):146–162.

M Haspelmath y A Sims. 2010. Understanding morphology oxford university press.

Wei He, Zhongjun He, Hua Wu, y Haifeng Wang. 2016. Improved neural machine translation with smt features. In *AAAI*. pages 151–157.

Gustav Herdan. 1966. *The advanced theory of language as choice and chance*. Springer-Verlag New York.

Karl Moritz Hermann y Phil Blunsom. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173* .

Karl Moritz Hermann y Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641* .

Geoffrey E Hinton y Sam T Roweis. 2003. Stochastic neighbor embedding. In *Advances in neural information processing systems*. pages 857–864.

Charles F Hockett. 1971. *Curso de lingüística moderna &d Charles F. Hockett*. Edit. Universitaria.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*. Association for Computational Linguistics, pages 29–32.

Ann Irvine. 2013. Statistical machine translation in low resource settings. In *HLT-NAACL*. pages 54–61.

Patrick Johansson y Miguel León-Portilla. 2010. *El español y el náhuatl: encuentros, desencuentros y reencuentros: discurso de ingreso a la Academia Mexicana de la Lengua 26 de agosto de 2010*. Universidad Nacional Autónoma de México.

Stig Johansson. 2007. Seeing through multilingual corpora. *Language and Computers* 62(1):51–71.

Katharina Kann, Manuel Mager, Ivan Meza-Ruiz, y Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. *arXiv preprint arXiv:1804.06024* .

Emmerich Kelih. 2010. The type-token relationship in slavic parallel texts. *Glottometrics* 20:1–11.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics* 21(3):223–245.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, y Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810* .

Alexandre Klementiev, Ivan Titov, y Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words .

Tomáš Kočiskỳ, Karl Moritz Hermann, y Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947* .

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn y Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*. Association for Computational Linguistics, pages 9–16.

Philipp Koehn, Franz Josef Och, y Daniel Marcu. 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*. Association for Computational Linguistics, pages 48–54.

Mikhail Kozhevnikov y Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *ACL (1)*. pages 1190–1200.

Hong-Seok Kwon, Hyeong-Won Seo, MA Cheon, y Jae-Hoon Kim. 2014. Iterative bilingual lexicon extraction from comparable corpora using a modified perceptron algorithm. *Journal of Contemporary Engineering Sciences* 7(24):1335–1343.

Luis Fernando Lara. 1979. *Investigaciones lingüísticas en lexicografía*, volume 89. Colegio de México, Centro de Estudios Lingüísticos y Literarios.

Adrien Lardilleux y Yves Lepage. 2009. Sampling-based multilingual alignment. In *Recent Advances in Natural Language Processing*. pages 214–218.

Adrien Lardilleux, Yves Lepage, y François Yvon. 2011. The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. *International Journal of Advanced Intelligence* 3(2):189–217.

Stanislas Lauly, Alex Boulanger, y Hugo Larochelle. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803* .

Michel Launey y Cristina Kraft. 1992. *Introducción a la lengua ya la literatura náhuatl*. Univ. Nacional Autónoma.

Angeliki Lazaridou, Georgiana Dinu, y Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. volume 1, pages 270–280.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, y Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *ACL (1)*. pages 1517–1526.

Robert B Lees y Noam Chomsky. 1957. Syntactic structures. *Language* 33(3 Part 1):375–408.

Omer Levy, Anders Søgaard, Yoav Goldberg, y Israel Ramat-Gan. 2016. A strong baseline for learning cross-lingual word embeddings from sentence alignments. *arXiv preprint arXiv:1608.05426* .

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pages 768–774.

Kevin Lund y Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28(2):203–208.

Minh-Thang Luong, Richard Socher, y Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. *CoNLL-2013* 104.

Laurens van der Maaten y Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Judit Martínez Magaz. 2003. El corpus paralelo: herramienta para el estudio de textos procedentes del inglés moderno temprano y sus traducciones al español .

Thomas Mayer, Bernhard Wälchli, Christian Rohrdantz, y Michael Hund. 2014. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. *Language Processing and Grammars. The role of functionally oriented computational models* pages 13–38.

Alfonso Medina-Urrea. 2000. Automatic discovery of affixes by means of a corpus: A catalog of spanish affixes. *Journal of quantitative linguistics* 7(2):97–114.

I Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. *arXiv preprint cmp-lg/9505044* .

Carlos-Francisco Méndez-Cruz, Alfonso Medina-Urrea, y Gerardo Sierra. 2016. Unsupervised morphological segmentation based on affixality measurements. *Pattern Recognition Letters* 84:127–133.

Tomas Mikolov, Kai Chen, Greg Corrado, y Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Tomas Mikolov, Quoc V Le, y Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, y Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, y Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, pages 880–889.

Paramita Mirza y Raffaella Bernardi. 2013. Ccg categories for distributional semantic models. In *RANLP*. pages 467–474.

David Mitchell. 2015. Type-token models: a comparative study. *Journal of Quantitative Linguistics* 22(1):1–21.

Ruslan Mitkov. 2005. *The Oxford handbook of computational linguistics*. Oxford University Press.

Christian Monson, Alon Lavie, Jaime Carbonell, y Lori Levin. 2004. Unsupervised induction of natural language morphology inflection classes. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*. Association for Computational Linguistics, pages 52–61.

Christian Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, y Alon Lavie. 2006. Building nlp systems for two resource-scarce indigenous languages: Mapudungun and quechua. *Strategies for developing machine translation for minority languages* page 15.

Robert C Moore. 2005. Association-based bilingual word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Association for Computational Linguistics, pages 1–8.

Denis Newman-Griffis y Eric Fosler-Lussier. 2017. Second-order word embeddings from nearest neighbor topological features. *arXiv preprint arXiv:1705.08488* .

Sonja Nießen y Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics* 30(2):181–204.

Nils J Nilsson. 2009. *The quest for artificial intelligence*. Cambridge University Press.

Douglas W Oard, David Doermann, Bonnie Dorr, Daqing He, Philip Resnik, Amy Weinberg, William Byrne, Sanjeev Khudanpur, David Yarowsky, Anton Leuski, et al. 2003. Desparately seeking cebuano. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*. Association for Computational Linguistics, pages 76–78.

Sebastian Padó y Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33(2):161–199.

Maria Pelevina, Nikolay Arefyev, Chris Biemann, y Alexander Panchenko. 2017. Making sense of word embeddings. *arXiv preprint arXiv:1708.03390* .

Jeffrey Pennington, Richard Socher, y Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

John R Pierce y John B Carroll. 1966. Language and machines: Computers in translation and linguistics .

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

Nima Pourdamghani, Marjan Ghazvininejad, y Kevin Knight. 2018. Using word vectors to improve word alignments for low resource machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. volume 2, pages 524–528.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 320–322.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, pages 519–526.

Philip Resnik y Noah A Smith. 2003. The web as a parallel corpus. *Computational Linguistics* 29(3):349–380.

Marta Ruiz Costa-Jussà, Parth Gupta, Rafael E Banchs, y Paolo Rosso. 2014. English-to-hindi system description for wmt 2014: deep source-context features for moses. In *ACL 2014 Ninth Workshop on statistical machine translation: proceedings of the workshop: June 26-27, 2014, Baltimore, Maryland, USA*. Association for Computational Linguistics, pages 79–83.

Geoffrey Sampson, David Gil, y Peter Trudgill. 2009. *Language complexity as an evolving variable*, volume 13. Oxford University Press.

Aquilino Sánchez. 2001. Investigación y análisis mediante corpus lingüísticos: el poder de atracción de las palabras. *Pathways of translation studies, Valladolid: Universidad* pages 11–46.

Charles Schafer y David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *proceedings of the 6th conference on Natural language learning-Volume 20*. Association for Computational Linguistics, pages 1–7.

Hinrich Schütze, Christopher D Manning, y Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press.

Hong-Seok Kwon Hyeong-Won Seo y Jae-Hoon Kim. 2013. Bilingual lexicon extraction via pivot language and word alignment tool. *ACL 2013* page 11.

C. E. Shannon. 2001. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.* 5(1):3–55. https://doi.org/10.1145/584091.584093.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, y Yuji Matsumoto. 2015. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pages 135–151.

Gerardo Sierra Martínez. 2017. *Introducción a los corpus lingüísticos*, volume 39. Instituto de Ingeniería, UNAM.

Anil Kumar Singh y Samar Husain. 2005. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and using Parallel texts*. Association for Computational Linguistics, pages 99–106.

Frank Smadja, Kathleen R McKeown, y Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics* 22(1):1–38.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

Samuel L Smith, David HP Turban, Steven Hamblin, y Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859* .

Radu Soricut y Franz Josef Och. 2015. Unsupervised morphology induction using word embeddings. In *HLT-NAACL*. pages 1627–1637.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28(1):11–21.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3):538–556.

Thelma D Sullivan y Miguel León-Portilla. 1976. *Compendio de la gramática náhuatl*, volume 18. Universidad nacional autónoma de México, Instituto de investigaciones históricas.

Kumiko Tanaka y Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 297–303.

Marc Thouvenot. 2011. Chachalaca en cen, juntamente. In *Compendio Enciclopedico del Nahuatl, DVD*. INAH.

Jörg Tiedemann. 2000. Word alignment step by step. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*. pages 216–227.

Jörg Tiedemann. 2003. Combining clues for word alignment. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 339–346.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*. pages 2214–2218.

Takashi Tsunakawa, Naoaki Okazaki, y Jun'ichi Tsujii. 2008. Building a bilingual lexicon using phrase-based statistical machine translation via a pivot language. In *COLING (Posters)*. pages 127–130.

Dan Tufiş y Ana-Maria Barbu. 2002. Lexical token alignment: Experiments, results and applications. In *Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas*. pages 458–465.

Dan Tufis, Radu Ion, Alexandru Ceausu, y Dan Stefanescu. 2006. Improved lexical alignment by combining multiple reified alignments. In *EACL*.

Peter D Turney y Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37:141–188.

Vladimir Naumovich Vapnik. 1998. *Statistical learning theory*, volume 1. Wiley New York.

Ljuba N Veselinova. 2006. *Suppletion in verb paradigms: bits and pieces of the puzzle*, volume 67. John Benjamins Publishing.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline .

Sami Virpioja, Ville T Turunen, Sebastian Spiegler, Oskar Kohonen, y Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *TAL* 52(2):45–90.

Martin Volk, Johannes Graën, y Elena Callegaro. 2014. Innovations in parallel corpus search tools. In *LREC*. pages 3172–3178.

Ivan Vulic y Anna-Leena Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings .

Warren Weaver. 1955. Translation. *Machine translation of languages* 14:15–23.

Dominic Widdows, Beate Dorow, y Chiu-Ki Chan. 2002. Using parallel corpora to enrich multilingual lexical resources. In *LREC*. pages 240–245.

Ludwig Wittgenstein. 2010. *Philosophical investigations*. John Wiley & Sons.

Hua Wu y Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation* 21(3):165–181.

Chao Xing, Dong Wang, Chao Liu, y Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1006–1011.

Barret Zoph, Deniz Yuret, Jonathan May, y Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201* .

Will Y Zou, Richard Socher, Daniel M Cer, y Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*. pages 1393–1398.