Recopilación de un corpus paralelo electrónico para una lengua minoritaria: el caso del español-náhuatl.

XIMENA GUTIÉRREZ VÁSQUES.
ELENA CAROLINA VILCHIS VARGAS.
ROCÍO CERBÓN YNCLÁN.
Universidad Nacional Autónoma de México.
Instituto de Ingeniería.
México, D.F.
xim@unam.mx
EVilchisV@iingen.unam.mx
RCerbonY@iingen.unam.mx

1. Introducción

El procesamiento del lenguaje natural (PLN), también llamado lingüística computacional, es un área multidisciplinaria que busca modelar el lenguaje humano desde una perspectiva computacional con el fin de generar tecnologías del lenguaje. El lenguaje humano es uno de los sistemas más complejos de los que tenemos conocimiento, lograr que una computadora imite funciones del lenguaje humano es una tarea titánica que requiere abordarse desde diversas áreas del conocimiento. Es por eso que en el PLN convergen la lingüística, la computación, la estadística, las matemáticas, entre otras áreas.

Entre las diversas tecnologías del lenguaje que usamos cotidianamente se encuentran los traductores automáticos, los asistentes de voz, los sistemas pregunta-respuesta, los buscadores, los sistemas de resumen automático por mencionar sólo algunos. Para generar muchas de estas tecnologías, es necesario dotar a los sistemas computacionales de grandes colecciones de documentos que les permitan encontrar patrones en el lenguaje humano. Estas colecciones de documentos son usualmente

conocidas como corpus y permiten el estudio del lenguaje humano así como el desarrollo de tecnologías.

En el presente trabajo abordaremos el tema de los corpus digitales y su uso en el PLN, en específico hablaremos de un tipo especial de corpus lingüístico llamado corpus paralelo. Los corpus paralelos son un recurso ampliamente utilizado para la creación de tecnologías de traducción. Nuestro trabajo consiste en la creación de un corpus digital paralelo español-náhuatl. Esperamos que nuestro caso de estudio sea de utilidad para aplicarse a otros pares de lenguas mexicanas y propiciar así la generación de tecnologías del lenguaje multilingües en nuestro país.

2. Corpus paralelos

Los corpus lingüísticos son un conjunto de documentos escritos y orales. Llisterri y Torruela en su artículo *Diseño de corpus orales y textuales* (1999) hablan de la necesidad de los corpus informáticos en las investigaciones humanísticas. Los autores afirman que en algunas áreas de las humanidades cada vez tienen más presencia los corpus, se hacen necesarios porque son una muestra de datos reales y permiten reproducir con máxima fidelidad las características del objeto de estudio esto implica recopilar en cantidades más o menos grandes muestras de la realidad que se quiere observar. Es importante tener en cuenta que, como se ha dicho, un corpus conforma un modelo de la realidad, pero debe ser lo suficientemente amplio para mostrar aspectos destacados y característicos, es relevante y mucho más útil tener un corpus grande, y lo que es más relevante de los corpus es que sean rentables, es decir, cuando ya se tiene conformado un corpus se pueda reutilizar para investigaciones posteriores y de cualquier área para la que se desee utilizar.

Hay muchos corpus importantes que son utilizados por estudiantes e investigadores los cuales se pueden consultar vía internet y son gratuitos. La idea de conformar un corpus parte de eso, de tener un objeto de estudio y explotar el potencial del corpus lo más que se pueda, desde cualquier análisis o incluso, crear tecnología de cualquier tipo.

Gracias a la informática, a los investigadores les es cada vez más fácil recopilar una vasta cantidad de datos, ordenarlos y tenerlos en formato electrónico para su consulta y análisis. El internet, por ejemplo, es un repositorio enorme con información actualizada al momento, con datos accesibles de usuarios de todo el mundo, la tendencia de recabar datos de la web para conformar un corpus es más común día a día.

Como podemos ver, el corpus lingüístico es un recurso ampliamente utilizado para estudiar fenómenos del lenguaje humano así como para desarrollar tecnologías del lenguaje. Dentro de los diferentes tipos de corpus textuales, se encuentran los corpus paralelos que son una serie de textos traducidos a una o varias lenguas, es decir, el texto original con su traducción en el mismo documento, la paralelidad de los textos puede ser desde nivel frase, párrafo, cuartilla o hasta por capítulo

El hecho de que los textos paralelos cubran los mismos significados y tengan idénticas funciones en ambas lenguas, los convierte en una gran fuente de información léxica bilingüe que permite el desarrollo de diversas herramientas del lenguaje. Los corpus paralelos son en la actualidad la materia prima esencial para entrenar a los sistemas de traducción estadísticos como Google Translate. Además permiten la creación de otros tipos de tecnología multilingüe, por ejemplo, son un recurso útil para la adquisición léxica multilingüe automática, permitiendo la construcción de diccionarios y otros recursos bilingües, por mencionar algunas aplicaciones.

Los corpus paralelos no son sólo útiles para entrenar sistemas informáticos, también resultan de utilidad para los traductores humanos que pueden utilizar corpus paralelos almacenados en línea como medio para encontrar cómo otros tradujeron ciertas palabras o frases dependiendo del contexto. Este tipo de herramientas derivadas de los corpus paralelos, por ejemplo las llamadas memorias de traducción, son recursos comúnmente usados por los traductores humanos.

3. Lenguas de bajos recursos digitales.

Los estudios en PLN se han concentrado en un subconjunto reducido de lenguas, ocasionando que muchas tecnologías del lenguaje estén solamente disponibles para lenguas como el inglés. Una lengua de bajos recursos digitales es aquella que no posee una vasta cantidad de textos publicados y disponibles de manera digital, puede ser porque no hay una cantidad grande de hablantes o no hay producción escrita.

Este tipo de lenguas representan un reto para el PLN precisamente por la escasez de datos: los métodos que funcionan razonablemente bien para otras lenguas necesitan replantearse y adaptarse al tratar con lenguas de bajos recursos digitales o poco estudiadas. Esto resulta particularmente importante cuando hablamos de corpus paralelos y creación de tecnologías multilingües. La calidad de muchas tecnologías del lenguaje es altamente dependiente del tamaño del corpus utilizado. Los sistemas de traducción automática estadística, por ejemplo, necesitan de corpus paralelos de cientos de millones de palabras para generar modelos estadísticos adecuados.

La existencia de la web ha detonado el fácil acceso a grandes cantidades de información de manera digital que permite la construcción de diversos corpus. Hoy en día, una fuente típica de corpus paralelos son las traducciones que se encuentran en la web, por

ejemplo, aquellas páginas web que publican su contenido en diversas lenguas como portales turísticos o gubernamentales.

Sin embargo, no todas las lenguas tienen la misma presencia en la web. Los datos arrojan que el mayor número de páginas están en inglés, en segundo lugar en ruso, seguido del alemán, posteriormente el japonés, etc (Wikipedia, 2014.) Las lenguas mexicanas como el náhuatl tienen una presencia casi nula en la web, lo cual dificulta recopilar un corpus mediante únicamente textos digitalizados encontrados en la web.

México es un país con una enorme diversidad lingüística, sin embargo, prácticamente no se ha desarrollado tecnología para las lenguas de las 68 agrupaciones lingüísticas que se hablan en el país. Esta situación convierte a nuestro país en un escenario ideal para desarrollar tecnologías del lenguaje y abordar los retos de la lingüística computacional que se presentan al trabajar con lenguas de bajos recursos digitales.

Nuestro caso de estudio se enfoca en el español-náhuatl. El náhuatl es una de las lenguas con mayor número de hablantes (1 millón 376 mil 26 hablantes), actualmente prevalece en más de 10 estados de la República mexicana es el náhuatl. A pesar de esto, se considera en riesgo de desaparición, hasta donde sabemos no existe un corpus paralelo para este par de lenguas digital y disponible para ser usado o descargado. El proyecto que se ilustra en el presente trabajo consistió en crear un corpus paralelo para el náhuatl que pueda ser accesible y que pueda ser de utilidad para propiciar la generación de aplicaciones así como estudios lingüísticos.

Esperamos que este esfuerzo se sume a otros para propiciar la conservación del importante patrimonio que representa la lengua. Como mencionamos anteriormente muy pocas tecnologías se han desarrollado para las lenguas mexicanas. Para el caso del náhuatl, nos gustaría destacar el esfuerzo que algunos investigadores en conjunto con el Instituto

Nacional de Antropología e Historia han realizado para generar tecnología para el náhuatl. Nos referimos al Compendio enciclopédico del náhuatl (CEN) el cual contiene cuatro grandes bases de datos del náhuatl:

- Gran diccionario del náhuatl (GDN) éste contiene palabras con variación dialectal diacrónica, es un compendio de distintos tipos de obras, diccionarios digitalizados. Contiene corpus paralelo porque estan las entradas en español y náhuatl de los diccionarios y contiene diferentes tipos de náhuatl y variantes contemporáneas.
- Editor de textos en náhuatl (TEMOA) es un gran corpus textual basado en códices.
- Diccionario de códices mexicanos (TLACHIA) es un diccionario de las imágenes que se encuentran en los códices.
- Analizador morfológico del náhuatl (CHACHALACA) debido a que el náhuatl es una lengua incorporante o aglutinante, este sistema descompone la palabra en sus partes y se puede analizar las diferentes interpretaciones de un gran número de palabras.

A continuación describiremos el proceso de recopilación y formación del corpus paralelo español-náhuatl así como las particularidades y dificultades presentadas al trabajar con este par de lenguas.

4. Corpus paralelo español - náhuatl

4.1 Recopilación.

En el artículo de Llisterri y Torruela se dan pautas para crear un corpus, una de ellas es decidir el tamaño del corpus, cuántos documentos, qué tipo de textos, de qué época y todo ello recae en la finalidad del corpus. En nuestro caso nos limitamos a que el corpus fuera paralelo y con cualquier tipo de náhuatl, de cualquier época, es decir que cualquier texto

encontrado con contenido paralelo estaba aprobado para conformar nuestro corpus, debido a la escasez de recursos bibliográficos.

La primera búsqueda bibliográfica con contenido paralelo se realizó dentro de la UNAM en las bibliotecas de diversos institutos tales como:

- Instituto de Investigaciones Filológicas
- Instituto de Investigaciones Históricas
- Instituto de Investigaciones Antropológicas
- Instituto de Investigaciones Bibliográficas
- Biblioteca Samuel Ramos de la Facultad de Filosofía y Letras
- Biblioteca Central

La segunda parte de la búsqueda fue en la biblioteca de la Escuela Nacional de Antropología e Historia. La recopilación requirió nuestra presencia en las bibliotecas para revisar que cada libro tuviera las especificaciones que buscábamos.

Todo lo anterior sumó un total de 33 libros (760 mil palabras aproximadamente.) La diversidad temática, cronológica y geográfica fue impresionante. Resultó un corpus ampliamente variado lo cual consideramos que sería un buen indicio, ya que mientras más rico, profuso y diverso fuera, tendría más capacidad de aprovechamiento.

Una vez obtenidos los libros realizamos una clasificación tentativa de carácter subjetivo para analizar los resultados de nuestra búsqueda. Decidimos clasificar nuestro corpus náhuatl-español de acuerdo con la época del náhuatl contenido en los textos, con género literario y de ser posible, región del país.

Nuestra clasificación arrojó 15 libros de lo que de manera medianamente arbitraria determinamos como náhuatl clásico y 15 libros de náhuatl actual, la mayoría con el

contenido paralelo en español contemporáneo, salvo unos cuantos de la época de la colonia que contienen español de aquellos tiempos.

Los géneros literarios a rasgos muy generales fueron de historia, costumbres, ciencia, teatro, poesía, narrativa, religiosos, recetarios y de carácter didáctico.

Algunas regiones fueron Milpa Alta, Puebla, Veracruz, Texcoco, Guerrero y la Huasteca.

4.2 Digitalización.

El segundo paso para la conformación del corpus fue el de convertir a formato electrónico todas las páginas de los libros. Esto, ciertamente, se logra a través de un escáner. El escáner digitaliza en una imagen cada página de los libros; estas imágenes se pueden guardar en PDF, este es un formato de almacenamiento de documentos digitales independiente de plataformas de software o hardware (Wikipedia, 2015.) Debido a que un archivo escaneado en PFD es una imagen, digamos como una fotografía, no bastó con tener los libros escaneados y en dicho formato. Por lo que un paso necesario fue el de procesar todas las páginas de cada libro en un tipo de software llamado OCR (reconocimiento óptico de caracteres) para este trabajo utilizamos un programa llamado ABBY Fine Reader, estos programas son sencillos: reconocen una letra o caracter de imagen y lo convierten en el mismo caracter pero en formato de texto como el de Office Word, Excel, o cualquier nota de texto, es decir, las imágenes las convierte a texto modificable. De esta manera es posible trabajar con el corpus, pues es material analizable por medio de diversos sistemas y así potencialmente crear tecnología con él.

A partir de este paso de conformación del corpus nos enfrentamos a diversos obstáculos que conllevó a una corrección y revisión manual de cada libro. Uno de los

primeros problemas fue el de lidiar con la escritura del náhuatl en los textos. El náhuatl es una lengua con una **e**scritura no normalizada, por ejemplo, el español está, casi en su totalidad, escrito con los mismos caracteres y los mismos acentos y tiene reglas ortográficas estipuladas, estará escrito de con las mismas pautas en cualquier lado, a diferencia del náhuatl que es una lengua con escritura bastante libre según el autor o la variante de la que se trate. La escritura del náhuatl no se atañe a ninguna regla y lidiar con esto fue el principal motivo de la corrección manual.

Otro problema fue enfrentarnos a estos programas OCR que no están entrenados para trabajar con dos lenguas en el mismo documento. Los OCR tienen predeterminados los idiomas e incluso son capaces de reconocerlo. Dentro del programa se encontraba el náhuatl, sin embargo la variante con la que trabaja el OCR es sólo una y muy prototípica, y nosotras contábamos con textos de múltiples variantes, por lo que decidimos leer todos los textos en español para facilitar la corrección, ya que con el español tenía un mejor desempeño y fallaba menos que cuando un texto en español lo leía en náhuatl.

Esta decisión tuvo repercusiones a lo largo de toda la corrección, debido a que muchos de los caracteres utilizados en algunos de los libros para el náhuatl no son empleados en español, sin embargo después de corregir algunos textos utilizando ambas opciones, concluimos que el programa cometía menos errores utilizando el español para ambos idiomas.

Ahora presentaremos unos ejemplos de los tipos de náhuatl para mostrar la diferencia de escritura:

- auh cuix noce çan ica yn iteyollotiliztin
- axka yĕl in másat umpa mukistik

- nečewa’ niawni tla’pias onkan
- aamo xi-chooca; xic- huiica n’éelootl n’tleen yi ootic-cuic

Como se ve en lo anterior, hay algunas marcas fonéticas que no están presentes en el español. En lingüística son llamados diacríticos, es decir, signos que equivalen a un sonido. Por ejemplo: ă, ĕ, ĭ, ŏ, ŭ, č, ç, etc. Debido a la que son grafías que no están presentes en el español, el programa ABBY OCR los suprimía en vez de conservarlos o los confundía por otras letras y cambiaba la palabra, por ejemplo, un error muy común era el de cambiar ŏ por d y muchísimas otras correcciones “falsas” que realizaba con estos signos porque evidentemente el programa suponía que era español.

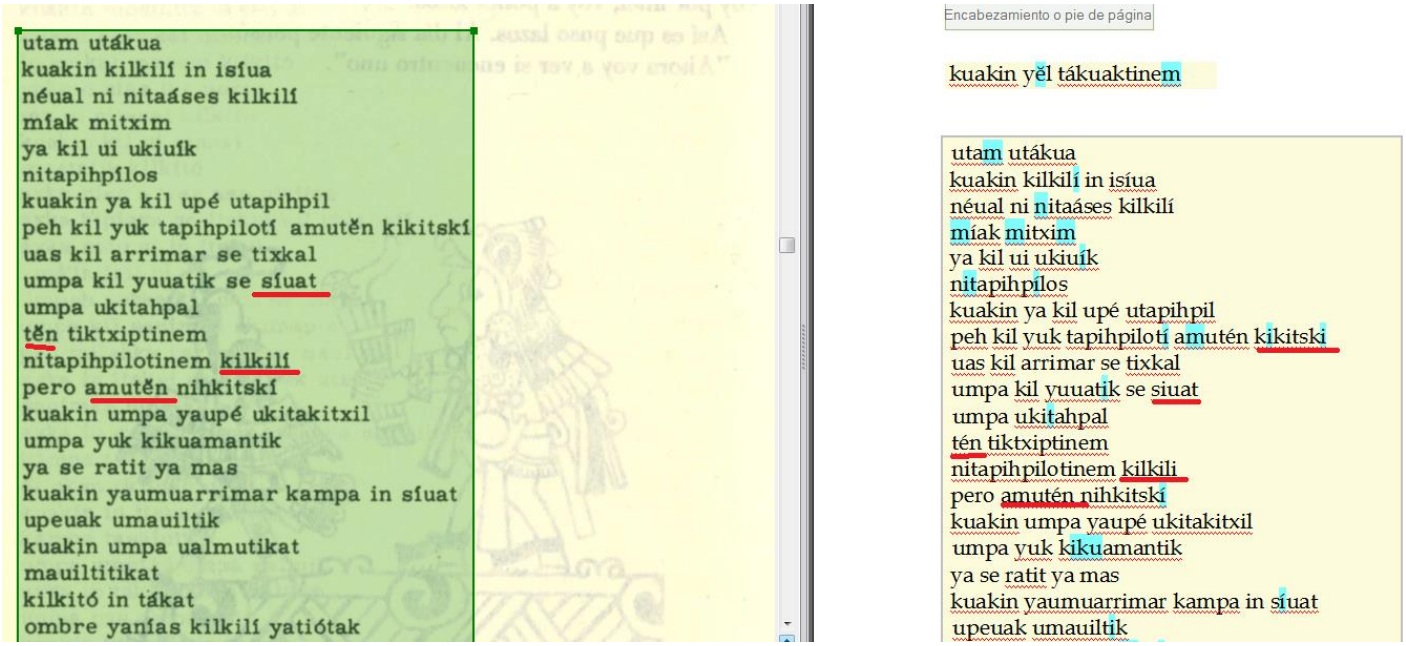


Figura 1. Ejemplo de una página en el OCR

Está de más decir que el náhuatl es una lengua completamente diferente al español y como se mencionó anteriormente, los documentos se leyeron en español por lo que el programa sobre-corregía las palabras por no ser patrones del español, frecuentemente realizaba correcciones asociativas**,** esto es que el programa trataba de encontrar las palabras

en náhuatl dentro del diccionario de palabras en español, de manera que cuando encontraba una parecida la cambiaba en vez de respetar sus letras originales y las modificaba para formar una palabra en español. Por ejemplo:

itla > ida,

ome > orne

ye > yo

Otro problema muy común fue la desaglutinación o separación de las palabras largas del náhuatl, ya que, como ya mencionamos, esta lengua es aglutinante, en la cual partículas se unen a una palabra para formar oraciones. Por ejemplo:

ocatea > oc - atea (palabra del español)

Bartolomé y otros nombres propios que no se acentúan en náhuatl los acentuaba porque en español sí se acentúan, pero reiteramos, todo esto se tuvo que corregir porque el programa leía los textos en español.

El siguiente obstáculo fue la tipografía o el tipo de letra de cada libro. Algunos libros tenían letras sin el contorno definido, un poco corridas o eran tipografías de alguna edición vieja así que la tinta ya estaba un poco desgastada. Estos errores eran comunes porque 33 libros tenían 33 tipografías diferentes junto con los antes mencionados diacríticos, conjuntamente se convirtió en un reto cada libro nuevo, cada parte en nàhuatl y cada parte en español. Ahora un ejemplo de una edición mal impresa:

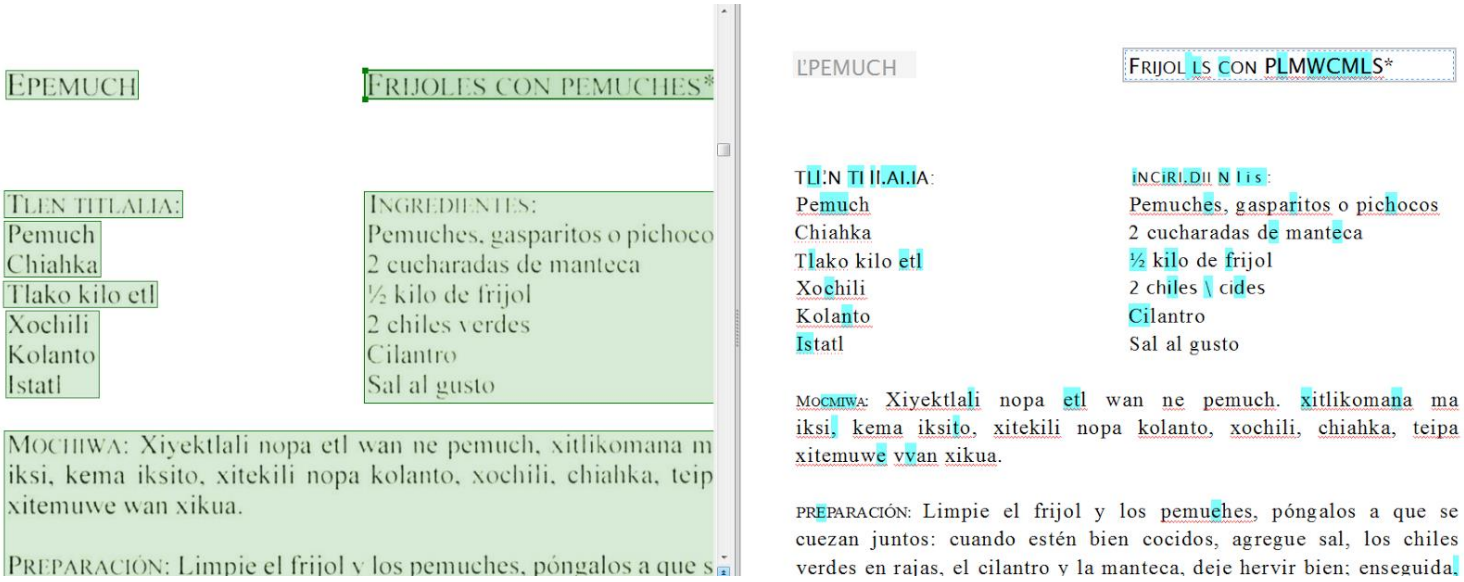


Figura 2. Ejemplo de una hoja con mala impresión.

De igual manera, encontramos numerosos libros cuyo texto compartía el espacio con imágenes, especialmente libros infantiles y didácticos.

A pesar de que ABBY tiene la función de distinguir entre imágenes y texto no siempre lo hace de manera correcta, suele confundir las imágenes con texto y viceversa.

Como vemos en la siguiente imagen ABBY coloca en un recuadro rojo lo que identifica como imagen y en verde el texto. Cuando comete errores como en este caso, debemos seleccionar manualmente las partes que debe leer como texto y eliminar las imágenes ya que para nuestro corpus no son útiles.

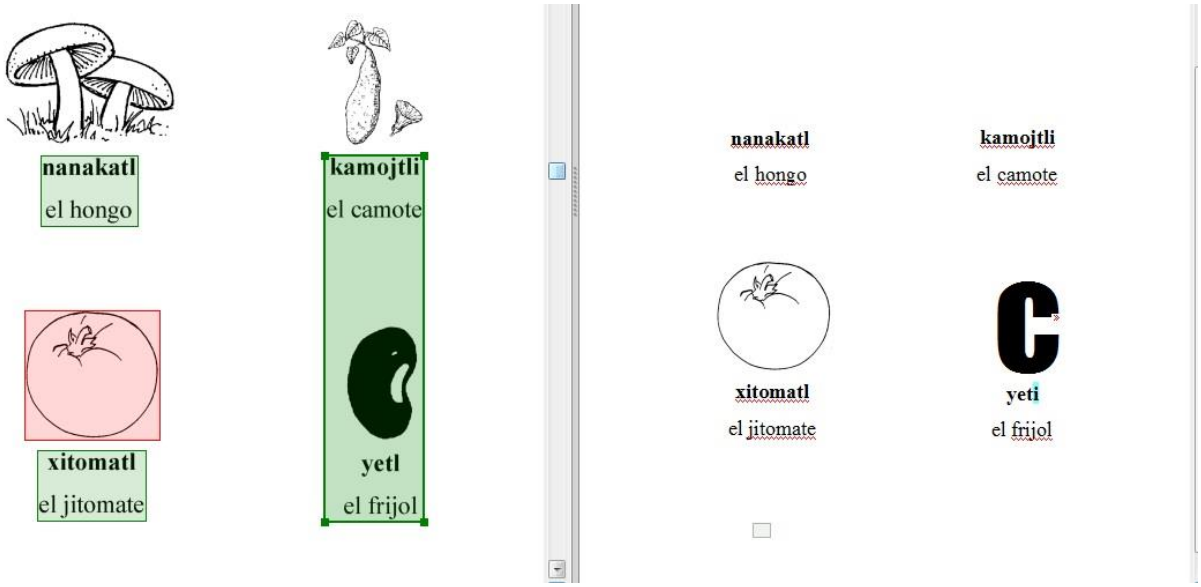


Figura 3. Un libro con texto e imagen en la misma página.

La situación se complica aún más cuando el espacio de texto e imagen no está tan bien delimitado como en el ejemplo anterior, esto es muy común en libros infantiles. La confusión del programa crece y es muy complicada la corrección manual, tal como se muestra en el siguiente ejemplo de *El viaje a Mictlan.*

Figura 4. Ejemplo de texto paralelo e imágenes en una misma página y cómo el OCR la reconocía.

Uno de los libros que conforman el corpus fue uno paralelo inglés-náhuatl. Pudimos notar que el OCR leía casi a la perfección el inglés, esto se debe a que es una lengua predominante en recursos electrónicos y además, este tipo de programas están principalmente diseñados para esta lengua.

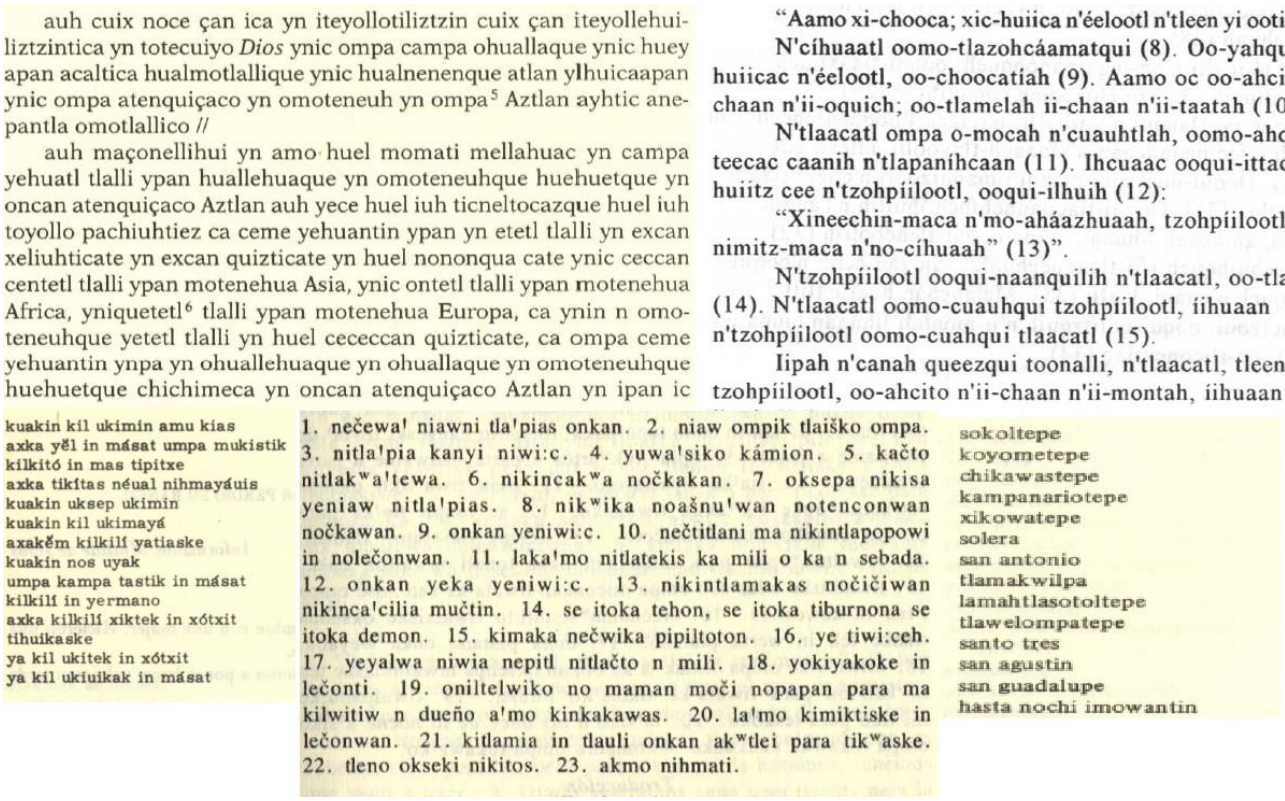
4.3 Corrección.

La principal estrategia de corrección a la que recurrimos fue la manual; buscábamos los errores del documento en el texto que nos daba ABBY mientras cotejábamos la imagen original en PDF. Esto era un trabajo arduo que requería de toda nuestra atención, sin embargo era poco eficiente ya que por las razones antes mencionadas eran demasiados errores y aunque a lo largo de corregir tantos textos fuimos adquiriendo experiencia para encontrar o saber en dónde era más probable que se equivocara el OCR, realizar toda la corrección de este modo nos hubiera llevado aún más tiempo del que tomó, por lo que implementamos dos estrategias más de corrección: enriquecer el diccionario de ABBY y la creación de un alfabeto nuevo, las cuales serán explicadas a continuación.

Nos encontramos con diversas palabras en las que cada vez que aparecían en el corpus, ABBY las cambiaba por alguno de los motivos anteriormente descritos. Cuando advertíamos que siempre cambiaba cierta palabra, podíamos agregarla, ya corregida, al diccionario con el que ABBY ya cuenta, de esta manera la próxima vez que hiciéramos una lectura con ABBY la escribiría correctamente. Es el caso de *Chalco* que invariablemente era corregida por *Chaleo,* al agregar la forma correcta al diccionario, ABBY ya no realizará

una falsa corrección nuevamente. Algunas veces esta estrategia fallaba cuando ambas formas, la original y la falsa corrección, existían, ya sea en náhuatl o en español, es el caso de *itla* invariablemente corregida por asociación a *ida*, en este caso no podíamos agregar *itla* al diccionario porque cuando apareciera *ida* sería corregida falsamente por la palabra en náhuatl.

La creación del nuevo alfabeto fue algo paulatino, en primer lugar debíamos tener en cuenta que como se dijo anteriormente, el náhuatl es una lengua no normalizada, es decir que no hay reglas ortográficas estipuladas, esto es que una misma palabra podía escribirse de muchas formas dentro de nuestros 33 libros, a esto debemos sumarle la gran diversidad de nuestro corpus tanto dialectal como diacrónica, como ya describimos, por lo tanto en el corpus hay gran variedad de escritura y de alfabetos, como podemos apreciar en la siguiente imagen:



auh cuix noce çan ica yn iteyollotiliztzin cuix çan iteyollehuiliztzintica yn totecuiyo *Dios* ynic ompa campa ohuallaque ynic huey apan acaltica hualmotlallique ynic hualnenenque atlan ylhuicaapan ynic ompa atenquiçaco yn omoteneuh yn ompa[5] Aztlan ayhtic anepantla omotlallico //

auh maçonellihui yn amo huel momati mellahuac yn campa yehuatl tlalli ypan huallehuaque yn omoteneuhque huehuetque yn oncan atenquiçaco Aztlan auh yece huel iuh ticneltocazque huel iuh toyollo pachiuhtiez ca ceme yehuantin ypan yn etetl tlalli yn excan xeliuhticate yn excan quizticate yn huel nononqua cate ynic ceccan centetl tlalli ypan motenehua Asia, ynic ontetl tlalli ypan motenehua Africa, yniquetetl[6] tlalli ypan motenehua Europa, ca ynin n omoteneuhque yetetl tlalli yn huel cececcan quizticate, ca ompa ceme yehuantin ynpa yn ohuallehuaque yn ohuallaque yn omoteneuhque huehuetque chichimeca yn oncan atenquiçaco Aztlan yn ipan ic

"Aamo xi-chooca; xic-huiica n'éelootl n'tleen yi ootic-cuic "(7).

N'cíhuaatl oomo-tlazohcáamatqui (8). Oo-yahqui, ooquihuiicac n'éelootl, oo-choocatíah (9). Aamo oc oo-ahcito n'iichaan n'ii-oquich; oo-tlamelah ii-chaan n'ii-taatah (10).

N'tlaacatl ompa o-mocah n'cuauhtlah, oomo-ahquettiteecac caanih n'tlapaníhcaan (11). Ihcuaac ooqui-ittac omkpa huiitz cee n'tzohpíilootl, ooqui-ilhuih (12):

"Xineechin-maca n'mo-aháazhuaah, tzohpíilootl. N'neh nimitz-maca n'no-cíhuaah" (13)".

N'tzohpíilootl ooqui-naanquilih n'tlaacatl, oo-tlapatlaqueh (14). N'tlaacatl oomo-cuauhqui tzohpíilootl, íihuaan n'tzohpíilootl oomo-cuahqui tlaacatl (15).

Iipah n'canah queezqui toonalli, n'tlaacatl, tleen oo-catca tzohpíilootl, oo-ahcito n'ii-chaan n'ii-montah, íihuaan oo-qui-

kuakin kil ukimin amu kias
axka yĕl in másat umpa mukistik
kilkitó in mas tipitxe
axka tikítas néual nihmayáuis
kuakin uksep ukimin
kuakin kil ukimayá
axakĕm kilkilí yatiaske
kuakin nos uyak
umpa kampa tastik in másat
kilkilí in yermano
axka kilkilí xiktek in xótxit tihuíkaske
ya kil ukitek in xótxit
ya kil ukiuikak in másat

1. nečewa' niawni tla'pias onkan. 2. niaw ompik tlaiško ompa. 3. nitla'pia kanyi niwi:c. 4. yuwa'siko kámion. 5. kačto nitlakʷa'tewa. 6. nikincakʷa nočkakan. 7. oksepa nikisa yeniaw nitla'pias. 8. nikʷika noašnu'wan notenconwan nočkawan. 9. onkan yeniwi:c. 10. nečtitlani ma nikintlapopowi in nolečonwan. 11. laka'mo nitlatekis ka mili o ka n sebada. 12. onkan yeka yeniwi:c. 13. nikintlamakas nočičiwan nikinca'cilia mučtin. 14. se itoka tehon, se itoka tiburnona se itoka demon. 15. kimaka nečwika pipiltoton. 16. ye tiwi:ceh. 17. yeyalwa niwia nepitl nitlačto n mili. 18. yokiyakoke in lečonti. 19. oniltelwiko no maman moči nopapan para ma kilwitiw n dueño a'mo kinkakawas. 20. la'mo kimiktiske in lečonwan. 21. kitlamia in tlauli onkan akʷtlei para tikʷaske. 22. tleno okseki nikitos. 23. akmo nihmati.

sokoltepe
koyometepe
chikawastepe
kampanariotepe
xikowatepe
solera
san antonio
tlamakwilpa
lamahtlasotoltepe
tlawelompatepe
santo tres
san agustin
san guadalupe
hasta nochi imowantin

Figura 5. Ejemplo de diferentes tipos de escrituras del náhuatl.

Como es posible imaginar cuando la escritura del náhuatl en cuestión usa los mismos caracteres que el español es más fácil para ABBY leerlos, caracteres que no contiene su alfabeto en español no puede leerlos y cometía gran cantidad de errores, por todo esto decidimos crear un nuevo alfabeto para aplicarlo a todas las lecturas que haría el OCR de los nuevos textos que tomó como base el de español y así cuando nos enfrentábamos a textos con distintos caracteres los agregábamos sucesivamente. Este alfabeto se fue enriqueciendo conforme avanzaba la corrección del corpus y cuando ABBY leía un texto cuyos caracteres ya habían sido agregados al nuevo alfabeto cometía considerablemente menos errores.

Como se pudo observar en lo descrito en esta sección, el procedimiento estándar de digitalización de un corpus que funciona bien para las lenguas con mayores recursos necesitó adaptarse al enfrentarnos a lenguas como el náhuatl. Durante el proceso de digitalización de los textos detectamos varios fenómenos interesantes entre los que se encuentran: Ultracorrecciones realizadas por el OCR, dificultad para lidiar con marcas fonéticas y tipografía, errores por el desgaste de las páginas, dificultad para reconocer textos con más de una lengua en la misma hoja. Estos retos podrían presentarse al digitalizar textos de otras lenguas mexicanas, esperamos que algunas de las soluciones adoptadas y presentadas en este trabajo puedan ser de utilidad para la formación de futuros corpus paralelos.

LIBROS QUE CONFORMAN EL CORPUS
Adivinanzas nahuas de hoy y siempre
Amerindia, leyendas y cantos nahuas.

Anales de Tepeteopan, De Xochitecuhtli a don Juan de San Juan Olhuatecatl.
Anales de Tlatelolco.
Antología del cuento náhuatl.
Augurios y abusiones.
Cantos indígenas de México.
Chimalpain Cuauhtlehuanitzin.Primera, segunda, cuarta, quinta y sexta relaciones de las differéntes histories originales.
Cuerpo humano e ideología, las concepciones de los antiguos nahuas.
Curso elemental de náhuatl clásico.
De Porfirio Díaz a Zapata, memoria náhuatl de Milpa Alta.
Documentos nauas de la Ciudad de México del siglo XVI.
Educación mexicana: antología de documentos sahaguntinos.
El anillo de Tlalocan
El náhuatl de Tetzcoco en la actualidad.
El viaje a Mictlán.
Historia de México narrada en náhuatl y español de acuerdo al calendario azteca.
La educación de los antiguos nahuas.
La llave del náhuatl.
La tierra nos escucha.
La tinta negra y roja, antología de poesía náhuatl.
La voz profunda. Antología de literatura mexicana en lenguas indígenas.
Lo que relatan de antes. Cuentos tének y nahuas de la Huasteca.
Los cuentos en náhuatl de Doña Luz Jiménez.
Método autodidáctico náhuatl-español.
Mitos y cuentos nahuas de la Sierra Madre Occidental.

Nahuatl as written.
Nican Mopouha
Primer Amoxtli Libro
Puede Hablar Náhuatl.
Recetario Nahua de Milpa Alta.
Recetario Nahua del norte de Veracruz.
Ritos, Sacerdotes y Atavíos de los Dioses.
Teatro Náhuatl.
Testimonios de la antigua palabra.
Trece poetas del Mundo Azteca.
Veinte himnos sacros de los nahuas.
Vida económica de Tenochtitlan.
Yancuitlalpan. Tradición y discurso ritual.

5. Utilización del corpus paralelo en las tecnologías del lenguaje

El corpus paralelo español-náhuatl antes descrito fue creado dentro del Grupo de Ingeniería Lingüística de la UNAM. Una de las finalidades del corpus es que esté disponible en línea para su consulta. Aunque el sistema aún se encuentra en desarrollo, la interfaz Web permite realizar búsquedas de palabras o frases en náhuatl o español, el resultado son aquellos fragmentos de texto paralelo que contienen la palabra buscada. Por ejemplo, en la Figura 6 se muestra el ejemplo de una búsqueda en el corpus paralelo, el usuario busca la palabra "mujeres" y el sistema regresa aquellos enunciados paralelos náhuatl-español que contienen la palabra buscada. La palabra buscada aparece resaltada al igual que su traducción correspondiente en el enunciado paralelo, en este caso "mujeres-cihuame". Esta última funcionalidad aparece en el prototipo con fines ilustrativos, sin embargo, la alineación

automática a nivel palabra para el español-náhuatl es un problema complejo que aún no está resuelto por completo.

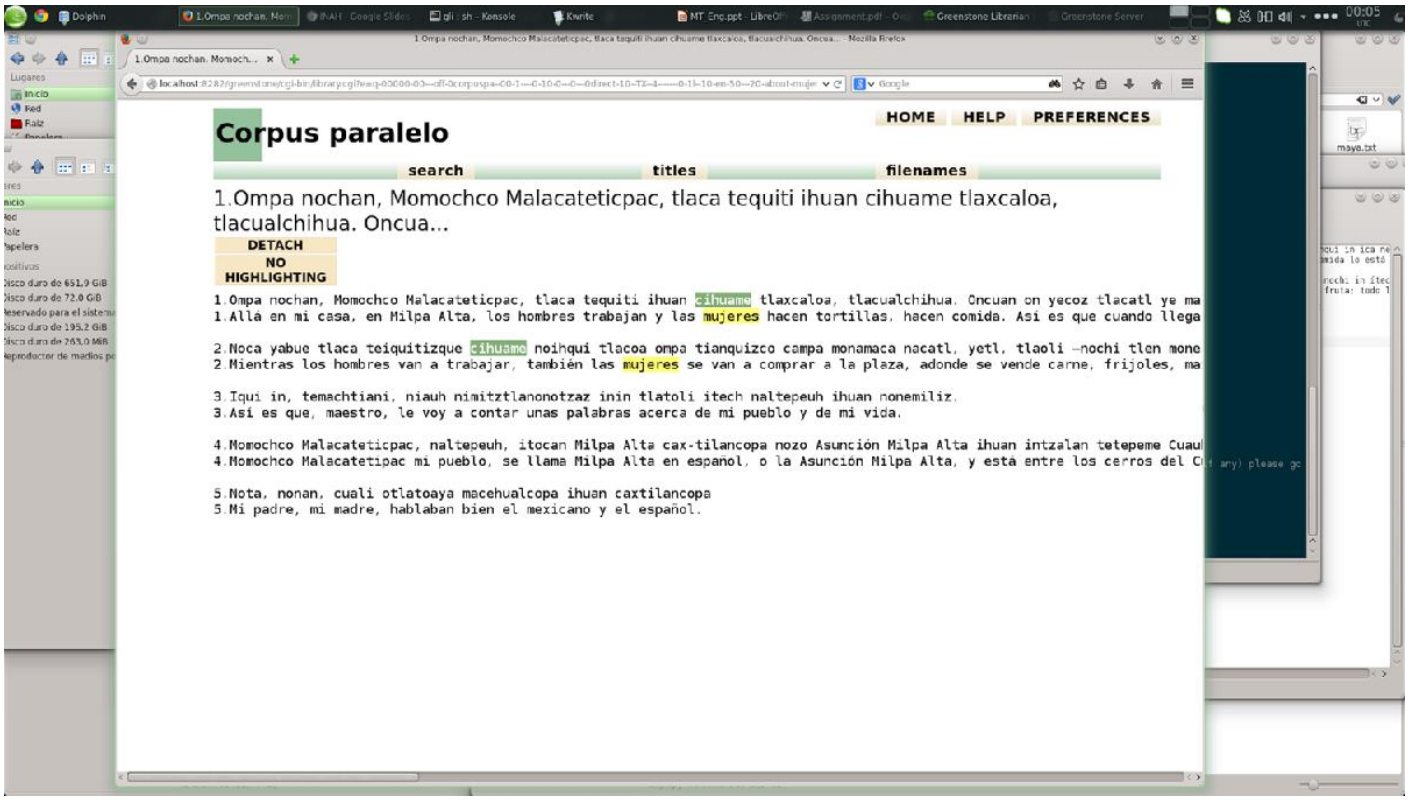


Figura 6. Esta imagen es un ejemplo de la interfaz web para corpus paralelo español-náhuatl.

El corpus paralelo español-náhuatl en línea puede ser utilizado con diferentes fines, por ejemplo, para realizar estudios lingüísticos de variación o de traducción. Por otra parte, el hecho de tener un compendio de textos paralelos para este par de lenguas es útil para la generación de tecnologías del lenguaje. En el Grupo de Ingeniería Lingüística se realiza actualmente una tesis doctoral de extracción léxica bilingüe automática a partir de este corpus paralelo, esto es, encontrar mediante métodos estadísticos correspondencias de

palabras o frases entre las dos lenguas para así construir, sin necesidad de un humano, un diccionario o lexicón bilingüe.

Aunque hablar de traducción automática entre español y náhuatl basada en el corpus antes descrito es un poco apresurado, la construcción de este corpus representa un primer paso para que en un futuro los traductores automáticos y otras tecnologías del lenguaje sean una realidad para las lenguas mexicanas.

La interfaz web del corpus paralelo aún no se encuentra disponible en línea, la versión final aún se encuentra en desarrollo además de que es necesario realizar un análisis sobre los derechos de autor aplicables a los textos antes de publicarlos en cualquier sitio web. En todo caso, el corpus paralelo español náhuatl será parte del sitio *www.corpus.unam.mx* que contiene diversos corpus creados dentro del Grupo de Ingeniería Lingüística de la UNAM.

A lo largo de este trabajo mostramos un panorama general sobre los corpus paralelos y como constituyen un recurso lingüístico valioso que permite el desarrollo de tecnologías de lenguaje, muchas de ellas enfocadas a la traducción automática. Trabajar con lenguas de bajos recursos digitales plantea interesantes desafíos en el desarrollo de tecnología, nuestro caso de estudio náhuatl-español puede ser aplicable a otras lenguas mexicanas para construir corpus digitales textuales y tecnologías del lenguaje

AGRADECIMIENTOS

Al Grupo de Ingeniería Lingüística (GIL) de la UNAM. Diversos académicos han brindado su valiosa asesoría en este proyecto, entre ellos, Dr. Gerardo Sierra Martínez, Dr. Alfonso Medina Urrea, Dr. Leopoldo Valiñas Coalla y Dr. Marc Thouvenot. También agradecemos

a Luis de la USI del Instituto de Ingeniería. Asimismo, un agradecimiento a Marco Antonio Lugo, Madai Ramirez, estudiantes de la carrera Lengua y Literaturas Hispánicas de la UNAM quienes ayudaron en la revisión y recopilación de textos del corpus paralelo.

Bibliografía

Brown, P. F., Lai, J. C., & Mercer, R. L. (1991, June). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistic*s (pp. 169-176). Association for Computational Linguistics.

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

Gómez Guinovart. Sacau Fontanela. *Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos*. http://www.sepln.org/revistaSEPLN/revista/33/33-Pag133.pdf

Koehn, P. (2005, September). Europarl: A parallel corpus for statistical machine translation. In *MT summit* (Vol. 5, pp. 79-86).

Instituto Nacional de Lenguas Indígenas. (INALI) Catálogo de las Lenguas Indígenas Nacionales, México. http://www.inali.gob.mx/clin-inali/

Torruela, J. Llisterri, J. (1990) "Diseño de corpus textuales y orales", en BLECUA, J.M.-Clavería, G. Sánchez, C. Torruela, J. (Eds) *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona: Seminario de Filología e Informática. Departamento de

Filología Española, Universidad Autónoma de Barcelona- Editorial Milenio. pp 45-77. Disponible en: http://liceu.uab.cat/~joaquim/publicacions/Torruella_Llisterri_99.pdf