2 Corpus lingüístico de la lengua maya de Yucatán: una propuesta metodológica

César Can Canul
Igor Vinogradov
Samuel Canul Yah
Alejandro Molina Villegas

1. Planteamiento del problema

En el siglo XXI la lingüística computacional ha expandido considerablemente su alcance. Distintas herramientas tecnológicas del procesamiento del lenguaje natural no solo se aplican a las lenguas mayoritarias con mayor difusión en el mundo y una tradición literaria bien establecida, sino también a las lenguas poco estudiadas que carecen de una gran cantidad de recursos informáticos. Ello abre un nuevo espacio para utilizar los métodos computacionales con el objetivo de recopilar y procesar información lingüística con relación a las lenguas indígenas y minoritarias, muchas de las cuales están en peligro de desaparición.

En el mundo moderno globalizado, el desplazamiento y la desaparición de lenguas se han acelerado, poniendo en riesgo la diversidad cultural del planeta. Se reconoce que la diversidad lingüística es esencial para la conservación del patrimonio de la humanidad, ya que cada lengua encarna la sabiduría cultural única de su comunidad de hablantes (UNESCO, 2003). De ahí surge la importancia de documentar, preservar y revitalizar las lenguas minoritarias (Himmelmann, 2006; Austin, 2016). México es uno de los países con mayor diversidad lingüística en el mundo, y por lo tanto los problemas del desplazamiento lingüístico y sus posibles consecuencias ocupan un lugar prominente en el contexto sociocultural mexicano, siendo objeto de varios estudios recientes de corte lingüístico y antropológico (Pellicer, 1999; León-Portilla, 2004; Valiñas, 2020; Nava, 2021).

La lingüística de corpus es una rama de la lingüística computacional que se dedica a la creación, organización, descripción y análisis de los corpus lingüísticos. A manera de una definición preliminar, se puede afirmar que un corpus lingüístico presenta una colección estructurada de textos digitales, debidamente seleccionados y anotados; véase Parodi (2008) para definiciones alternativas. Esta colección de textos pretende servir para objetivos específicos de investigación lingüística; por ejemplo, para realizar una descripción teórica de un dominio léxico o gramatical de la lengua o como un instrumento que permite verificar hipótesis acerca de ciertos fenómenos que se presentan en la lengua (Crystal, 1991; Leech, 1991).

Los recientes estudios multidisciplinarios han demostrado que la lingüística computacional, el procesamiento del lenguaje natural y la lingüística de corpus pueden ir de la mano con la documentación y revitalización lingüística, actuando a favor de la preservación de la diversidad lingüística y cultural (Rojas, 2017; Mager *et al.*, 2018). Se ha corroborado que la lingüística de corpus y la documentación de lenguas minoritarias se pueden apoyar mutuamente (Vinogradov, 2016). Un corpus lingüístico es una herramienta adecuada y eficaz para realizar la documentación lingüística. En México, la aplicación de la lingüística de corpus a lenguas minorizadas forma una línea de investigación que todavía está en crecimiento (Reyes y García, 2021; Angel *et al.*, 2021; Sierra, Bel y Mota, 2021). El objetivo principal de este artículo es presentar las múltiples ventajas que puede tener un estudio de corpus para los estudios de lenguas poco descritas. De esta manera se busca incentivar la creación de corpus lingüísticos para distintas lenguas indígenas de México.

2. Consideraciones culturales

La lengua maya de la península de Yucatán actualmente cuenta con más de 770 mil hablantes (INEGI, 2020). Es la lengua indígena más hablada en el territorio mexicano, distribuida en los Estados de Yucatán, Quintana Roo y Campeche, y una parte de Belice. Can y Gutiérrez (2016, p. 11) señalan que es incorrecto pensar que el náhuatl es la lengua indígena de México con el mayor número de hablantes, ya que el náhuatl no es una lengua, sino un subgrupo conformado por diferentes lenguas. En cambio, el maya yucateco sí es una lengua unitaria.

A pesar de esta difusión relativamente amplia, la continuidad y existencia de la lengua maya se encuentran en riesgo ante los procesos actuales de globalización (Arzápalo y Gubler, 1997; Sobrino y Paz, 2008). Desafortunadamente, el desplazamiento lingüístico y la pérdida que atraviesa la lengua y cultura maya se han acelerado considerablemente en las últimas décadas (Pfeiler, 1997, 1998; Sánchez, 2009). La transmisión intergeneracional, siendo la columna principal que sostiene la vitalidad de una lengua, se encuentra fracturada gravemente (Sima, Perales y Be, 2014). Es decir, muchos padres y madres mayahablantes ya no les están enseñando su lengua y cultura a sus hijos e hijas.

Los motivos para ello son diversos e incluyen la discriminación sistemática, la nula existencia de políticas lingüísticas y educativas favorables para el fortalecimiento lingüístico y cultural, las actitudes negativas hacia la lengua de los no hablantes y de los propios hablantes, la falta de recursos educativos y tecnológicos, entre otros factores; véase Briceño (2017) para una discusión más detallada. La ausencia de la tradición escrita y las normas de ortografía universalmente aceptables y reconocidas dentro de las comunidades también afecta de manera negativa al uso de la lengua maya en el ámbito educativo (Guerrettaz, Johnson y Ernst-Slavit, 2020). Como consecuencia directa de estos procesos sociolingüísticos, los niños no adquieren la lengua maya o no la utilizan en su vida diaria.

Es evidente que si no se toman acciones e iniciativas que reviertan este proceso de desplazamiento, en algún momento no muy lejano los hablantes de la lengua maya de Yucatán estarán condenados a ver morir su lengua, y con ella su cultura, saberes y conocimientos ancestrales que han sobrevivido hasta ahora, de generación a generación, gracias a la transmisión intergeneracional. La

situación sociocultural ha propiciado que las comunidades vayan dejando a un lado ciertos saberes, aunque también han surgido nuevos, debido a que la misma gente ha ido adaptando su lengua para poder describir su entorno y la relación que mantiene con el mismo (Villanueva, 2008). Algunas acciones a favor de la difusión, fortalecimiento, sensibilización y transmisión de la lengua y cultura maya a las nuevas generaciones se han venido desarrollando desde hace tiempo. Sin embargo, apenas recientemente se le reconoce de manera oficial como patrimonio cultural inmaterial de Yucatán mediante el decreto 474/2022 publicado el 21 de marzo del año 2022 en el Diario Oficial del Estado de Yucatán.

En este contexto el corpus lingüístico de la lengua maya yucateca proporcionará una herramienta innovadora que pretende contribuir a la salvaguarda del patrimonio cultural inmaterial y lingüístico. Mediante la documentación oral en archivos de audio y video se llevará a cabo el registro del habla cotidiana sobre diferentes temáticas y contextos sociolingüísticos, de modo que su legado se resguarde para ésta y las siguientes generaciones.

3. Consideraciones lingüísticas

El corpus lingüístico puede ser definido como "un conjunto de textos de materiales escritos y/o hablados, debidamente recopilados para realizar ciertos análisis lingüísticos" (Sierra, 2017, p. 4). A mediados del siglo XX, cuando la lingüística de corpus estaba todavía por aparecer, el procesamiento de estos textos se realizaba con fichas de papel que en nuestra época fueron reemplazadas por archivos digitales y bases de datos computacionales. La presentación digital de datos lingüísticos ofrece varias ventajas: se agiliza la búsqueda, se puede almacenar un volumen grande de textos, se pueden trasladar, conservar y compartir los repositorios electrónicos de manera fácil y cómoda, entre otras ventajas.

Existe una gran variedad de corpus según el alcance, el contenido, los criterios que se utilizan para seleccionar materiales, el propósito, etc. Dentro de estas posibilidades tan diversas, nuestro interés se enfoca solamente en los audios y videos con contenido de lengua oral. Es decir, nuestro propósito es crear un corpus de archivos digitales para la comunidad de hablantes de la lengua maya yucateca, donde se sienta representada a través de sus formas particulares de habla cotidiana. Además, se pretende que el corpus resguarde los conocimientos y saberes tradicionales de las personas mayahablantes, manifestados en la lengua oral.

Por esta razón optamos por excluir del corpus los textos escritos, tanto periodísticos actuales (por ejemplo, las traducciones de los artículos de *La Jornada Maya*), como las fuentes publicadas en los años anteriores (por ejemplo, la colección de cuentos de Andrade y Máas Collí, 1999). De esta manera queremos garantizar que el corpus sea un modelo fiel de la realidad lingüística que se observa en la península de Yucatán.

Dos aspectos son de suma importancia en el diseño de cualquier corpus lingüístico: la representatividad y el equilibrio. Como ningún corpus es capaz de incluir todos los actos lingüísticos existentes, la selección de textos debe reflejar el comportamiento lingüístico de toda la comunidad de hablantes en general. Es decir, los textos o grabaciones en el corpus deben representar todos los

dialectos, los géneros de discurso y los perfiles sociolingüísticos de los hablantes, entre otras variables. Además, para satisfacer el criterio de equilibrio, la cantidad de textos que cumplen con estas características debe ser suficiente y, en la medida de lo posible, coincidir con lo que se da en la realidad. En otras palabras, para que el corpus pueda ser considerado una muestra representativa y equilibrada, la cantidad de textos que representan cada una de las variables sociolingüísticas necesariamente debe coincidir con el respectivo porcentaje de la población mayahablante.

Los corpus lingüísticos de lenguas minoritarias y poco estudiadas por lo común no son representativos ni equilibrados, a consecuencia de que se crean a la par con las actividades de documentación lingüística (Vinogradov, 2016). Son corpus que se elaboran con base en los materiales textuales escritos u orales recopilados por un investigador individual o un grupo de investigadores relativamente pequeño en el transcurso de su trabajo de campo con hablantes de la lengua. Estos corpus suelen incluir todos los materiales disponibles para garantizar el mayor tamaño posible del corpus, sin la preocupación por la representatividad y el equilibrio de la muestra. Sin embargo, la idea de nuestro proyecto es elaborar un corpus representativo y equilibrado, a pesar de las obvias dificultades al momento de recopilar y seleccionar materiales.

El andamiaje digital de la lengua maya se construirá mediante tres actividades específicas: 1) documentación lingüística sistematizada en forma de grabaciones de audio y video; 2) transcripción de estas grabaciones y anotación morfológica de las transcripciones; y 3) creación de un repositorio electrónico. El acceso a estos recursos se proveerá a través de una plataforma digital. Las primeras dos actividades se describen a continuación, mientras que los detalles de la tercera actividad, que estará a cargo de los ingenieros de software, se discutirán en la siguiente sección.

Actividad específica 1. La documentación lingüística sistematizada de la lengua maya en la península de Yucatán

Para llevar a cabo la documentación lingüística, primero se elaborará un mapeo de las variantes de la lengua maya con base en los estudios preliminares (Blaha Pfeiler y Hofling, 2006; Hernández, 2019). Se definirá la cantidad de variantes que componen la lengua y las regiones donde se ubican. Para esta labor, se contará con la participación de investigadores expertos en el estudio del tema de variación dialectal de la lengua maya.

Se realizarán estudios sociolingüísticos para diseñar la estructura del contenido del corpus, contemplando los principios fundamentales de representatividad y equilibrio. Se utilizará la información demográfica y estadística para definir la cantidad de textos que debe ser recopilada en cada localidad y para cada variante lingüística. Se determinarán las características sociolingüísticas de las personas que serán grabadas, cuidando que los archivos audiovisuales sean diversos y equilibrados, es decir, con representatividad de género, diversos grupos de edad, grados de estudio, actividades económicas, entre otros aspectos.

La participación activa de la comunidad mayahablante es esencial para llevar a cabo la actividad de documentación. Mediante una convocatoria se conformará un equipo de 10 personas mayahablantes nativos pertenecientes a distintas regiones dialectales previamente identificadas. Al equipo de

documentadores, se les concientizará sobre el valor lingüístico y cultural de su lengua y se les capacitará en la lecto-escritura de la lengua maya. Este paso es necesario, ya que lamentablemente el porcentaje de hablantes alfabetizados es muy bajo, debido a que en el ámbito educativo, la alfabetización se da históricamente en español y, además, no existen suficientes recursos escritos para fomentar la lectoescritura. Asimismo, se les capacitará en el manejo de equipo de audio y video, así como en técnicas de documentación lingüística.

Una vez concluida la capacitación, el equipo de documentación empezará a recopilar material audiovisual de buena calidad y directamente de la comunidad, de algún registro oral de práctica comunicativa, como puede ser un cuento, una leyenda o una anécdota. También se registrarán prácticas comunicativas que involucren la interacción de dos o más personas como puede ser una plática cotidiana, una ceremonia, una asamblea comunitaria, algún oficio, etc. Cada archivo estará acompañado de una lista de metadatos que serán necesarios para poder identificar y clasificar la grabación dentro del corpus. Los metadatos incluirán el lugar, la fecha, el género y edad del hablante, el equipo de grabación, entre otra información. La documentación audiovisual será el insumo para el funcionamiento óptimo del corpus maya.

Actividad específica 2. La transcripción de las grabaciones y la anotación morfológica

La transcripción y la anotación lingüística es la parte medular del proyecto, ya que implica todo el levantamiento y procesamiento de la información que se obtuvo mediante las actividades de documentación. Se espera que los miembros del equipo de documentación logren obtener el conocimiento necesario de lectoescritura para poder participar en la actividad de transcripción.

Las grabaciones se transcribirán y se anotarán en el programa ELAN que es un software informático de libre acceso desarrollado por el Instituto Max Planck. Es el instrumento más adecuado y más difundido en la comunidad académica de lingüistas que se dedican a la documentación de lenguas en peligro de extinción (Brugman y Russel, 2004; Crasborn y Sloetjes, 2014). La anotación es lo que da más valor al corpus, ya que es el elemento que genera las distintas posibilidades de análisis y búsqueda. Por el momento, se consideran las siguientes capas o niveles de anotaciones:

1. Transcripción libre
2. Transcripción ortográfica (fonémica)
3. Segmentación morfológica
4. Glosado
5. Traducción libre en español

La transcripción libre se realizará con los caracteres del alfabeto actual de la lengua maya, buscando acercarse a una transcripción fonética y representar la manera particular de hablar en la medida de lo posible. Hacer transcripción mediante el alfabeto fonético internacional es un trabajo que requiere conocimientos profundos y horas de práctica especializada, lo que no puede ser cubierto con solamente una breve capacitación. Estamos conscientes de que es una simplificación; sin embargo, por cuestiones de tiempo es más viable preferir el uso de los caracteres del alfabeto actual que los símbolos fonéticos.

Para la transcripción ortográfica se seguirá lo descrito en las normas de la escritura de la lengua maya (Briceño y Can Tec, 2014). Estas convenciones ortográficas son recientes – ver Brody (2004) para los

antecedentes – y todavía no han logrado una aceptación universal en la comunidad mayahablante. Además, varios problemas de escritura no se mencionan o no se resuelven en dicha norma; véase Lehmann (2018) para la crítica de algunas propuestas. Como la tradición ortográfica apenas se está estableciendo actualmente en la península de Yucatán, todos los casos problemáticos se discutirán en equipo, con el objetivo de sistematizar la transcripción ortográfica. En un futuro, estas reflexiones podrán contribuir a la revisión de las normas existentes de la escritura y a la elaboración de propuestas concretas para su mejora.

En cuanto a la segmentación morfológica y al glosado, estarán a cargo de dos especialistas en el tema. Se contratará a dos expertos en la morfología del maya yucateco que se encarguen específicamente de estas dos capas de anotación. La segmentación morfológica indicará la estructura interna de las palabras, separando los afijos y clíticos de las raíces. La capa del glosado adscribirá a cada morfema su significado léxico y/o gramatical. La misma capa va a llevar anotaciones de otro tipo que serán utilizadas para ampliar las opciones de búsqueda; por ejemplo, las clases de palabras (sustantivos, adjetivos, verbos, preposiciones, etc.), los préstamos, los nombres propios, etc.

La traducción libre al español es opcional, ya que el corpus de la lengua maya no pretende ser bilingüe ni paralelo, aunque para varias tareas del procesamiento de lenguaje natural podría ser muy útil contar con una traducción aproximada. Para cada texto individual, el anotador podrá decidir si incluir la traducción o no. Un fragmento de texto en lengua maya anotado en el programa ELAN aparece en la Figura 1. Se incluyen cuatro capas de anotación: la transcripción fonémica, la segmentación morfológica (con símbolos "=" para clíticos y "-" para afijos), el glosado y la traducción al español.

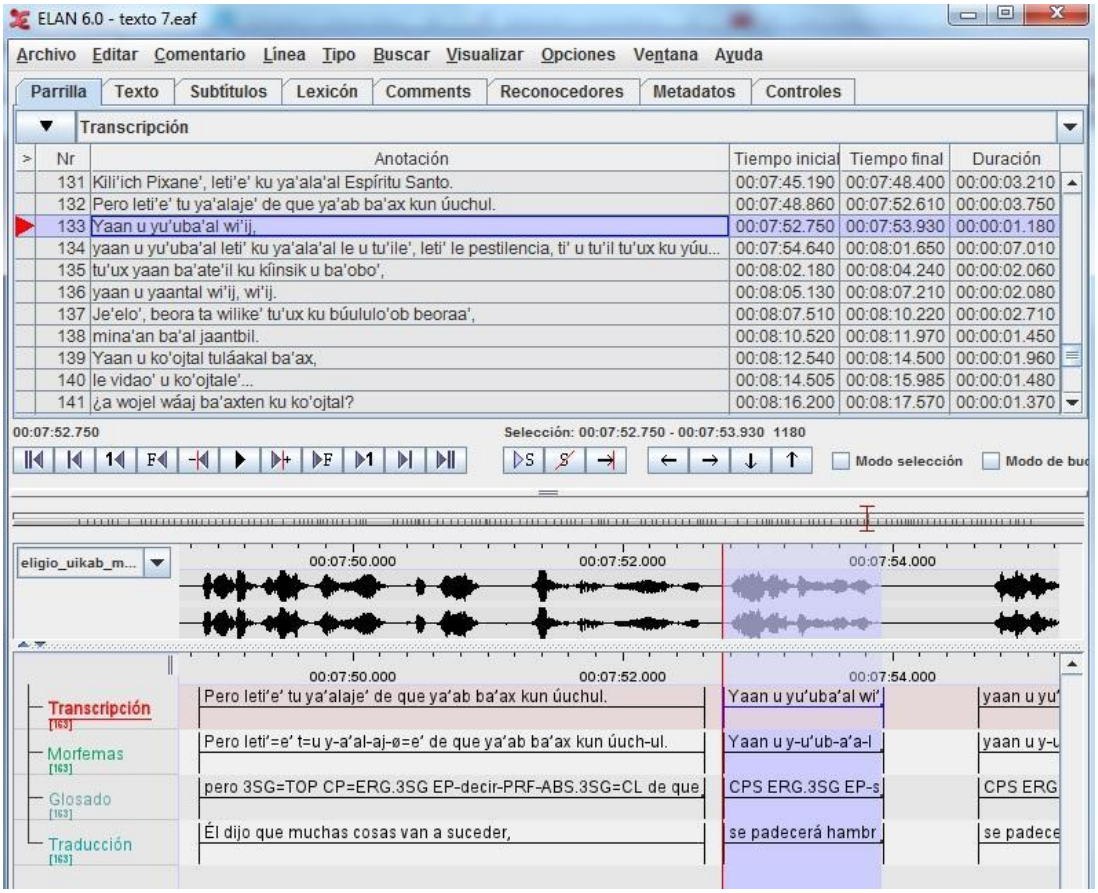

Figura 1. Las capas de anotación lingüística en el programa ELAN.

4. Descripción de la plataforma digital del corpus maya

Como el corpus de la lengua maya de Yucatán es un producto electrónico, es necesario implementar una plataforma digital para presentar los datos y crear una interfaz para la interacción con los usuarios. Esta plataforma, además de proveer el espacio para almacenar los archivos de audio, la base de metadatos y las anotaciones lingüísticas, incluirá el mecanismo de búsqueda en el corpus.

El advenimiento de las plataformas digitales a principios de siglo sacudió a la sociedad cambiando por completo los mecanismos de comunicación humana. Como consecuencia, las lenguas vivas están más o menos representadas en el ciberespacio según la cantidad de datos disponibles que de ellas existan y las plataformas digitales que los soportan (Van Dijck, 2013). Wikipedia por ejemplo está disponible en 325 idiomas pero al ordenar las distintas versiones por número de artículos resulta que la lengua mejor representada es el inglés (con más de 6 millones de artículos) y la peor representada es el kanuri, un idioma nilo-sahariano hablado en Nigeria, Níger y Chad, para el cual hay solamente un artículo.[1] Como consecuencia directa, mucha de la tecnología actual, como búsquedas por palabra, corrección ortográfica, traducción automática, extracción terminológica y resumen automático, se consolidan para las lenguas representadas y son inexistentes o incipientes para las lenguas poco representadas. Ante este panorama, es importante generar datos y recursos informáticos para el maya.

Como parte del proyecto del corpus, un equipo conformado por investigadores, tecnólogos y programadores profesionales desarrollará módulos de procesamiento de lenguaje adaptados para el maya. Dicho desarrollo incluye, en gran medida, trabajo de investigación puesto que no hay antecedentes en el estado del arte que propongan herramientas de procesamiento de lenguaje natural para este idioma. Por tanto, se deberán diseñar y codificar nuevos módulos de software especializados en procesamiento de maya creando un antecedente importante de cara al desarrollo tecnológico en esta lengua. Los módulos de software y otros recursos informáticos integrarán la plataforma digital del corpus maya, un sistema de recuperación de información online que permitirá realizar búsquedas por palabras clave en documentos audiovisuales en maya, así como visualización y descarga de datos.

Una vez liberada una versión estable de la plataforma digital del corpus maya, se elaborarán los manuales de usuario y manuales técnicos correspondientes. Se integrará el código fuente en repositorios en la nube de manera tal que al término del proyecto se tendrá la posibilidad no solo de dar continuidad al desarrollo de la plataforma sino de incrementar los recursos informáticos para la lengua maya y los conocimientos del procesamiento de ésta y otras lenguas minorizadas.

Entre los resultados esperados al final del desarrollo de la plataforma se encuentran los siguientes.

- Por lo menos 3 módulos (podrían ser más) de software especializados para procesamiento de la lengua maya.

[1] Datos extraídos de https://meta.wikimedia.org/wiki/List_of_Wikipedias/es visitado el 25 de marzo 2022.

Preprocesamiento: algoritmos para homogeneización de texto que producen versiones óptimas para procesos ulteriores de cara a la recuperación de información. Incluye: manejo de diacríticos, capitalización, tokenización y lematización. La tokenización es un proceso para separar palabras (unidades de una lengua) a partir de un texto de corrido. La lematización es un algoritmo para determinar si dos palabras tienen la misma raíz a pesar de sus diferencias superficiales (Jurafsky y Martin, 2000).

Indexación: algoritmos para la generación de estructuras de datos e índices que permiten almacenar y vincular información de manera organizada y optimizada en un sistema de recuperación de información. El módulo incluirá generación de bolsas de palabras, generación de n-gramas y vectorización (ver Manning, Raghavan y Schütze, 2010).

Consulta: algoritmos que establecen el funcionamiento de los sistemas de recuperación a partir de una barra de búsqueda. Incluye: procesamiento de queries y operadores de búsqueda.

- Una versión estable de la plataforma virtual de la lengua maya disponible para consulta y para generar productos que contribuyan a la revitalización de la lengua. Incluye consulta en la plataforma y descarga de datos.
- Sistematización del proceso de la plataforma para contribuir a su sustentabilidad y a la generación de otros corpus lingüísticos a través de repositorios de código fuente abiertos, manuales técnicos y manuales de usuario también de acceso abierto.

- Invitación a la participación de autoridades y actores clave sensibilizados en torno a la importancia y el potencial de uso de un corpus lingüístico maya a través de talleres de capacitación, manuales en línea para el uso del corpus lingüístico maya y propuestas de proyectos de tesis e investigación vinculada al proyecto.

5. A manera de conclusión

La creación de un corpus lingüístico implica un trabajo colaborativo. El éxito del proyecto del corpus de la lengua maya va a depender de la participación de, al menos, tres partes: los lingüistas, los activistas nativohablantes y los programadores. La contribución de los lingüistas es el diseño del corpus (contemplando los principios de representatividad y equilibrio) y la anotación de textos. La parte esencial es la capacitación de hablantes nativos en cuanto a la grabación de audio y video en comunidades indígenas, transcripción de lengua hablada, y el manejo del programa ELAN. El desarrollo de la base de datos, la interfaz del corpus y los algoritmos de búsqueda estarán a cargo de los ingenieros de software.

El corpus de la lengua maya de Yucatán es pensado como un corpus de referencia; es decir, un modelo representativo y equilibrado de toda la lengua oral de la península. Estas dos características aproximan el corpus de la lengua maya a los corpus de las grandes lenguas europeas como el español o el inglés que cuentan con miles de millones palabras y que dieron inicio a la lingüística de corpus en el siglo pasado. Este proyecto pretende demostrar que los pequeños corpus construidos con base en materiales

de documentación lingüística pueden tener mucho en común con los grandes corpus de lenguas mayoritarias.

A pesar de que el maya yucateco es la lengua más estudiada de toda la familia maya, las gramáticas de referencia del maya yucateco todavía no existen. Ninguna de las descripciones gramaticales (por ejemplo, Andrade, 1955) tiene alcance universal y abarca todas las reglas y construcciones de la lengua. En este sentido, la creación de un corpus de referencia es aún más importante, ya que puede llegar a ser un instrumento clave para la elaboración de una gramática de referencia en un futuro.

El impacto principal del corpus de la lengua maya para la sociedad es la contribución a la revitalización de la lengua y cultura maya en las comunidades mayahablantes de la península de Yucatán. Adicionalmente, el corpus es un valioso recurso lingüístico, cultural y tecnológico, del cual pueden generarse diversos recursos educativos, tecnológicos y de ocio, encaminados hacia el fortalecimiento de la lengua maya. El corpus de la lengua maya va a ser el primer proyecto de este tipo y alcance para una lengua indígena de México.

Referencias

Andrade, M. J. (1955). *A grammar of modern Yucatec*. University of Chicago Library.

Andrade, M. y Máas Colli, H. (Eds.). (1999). *Cuentos mayas yucatecos*. Tomo I. Universidad Autonoma de Yucatán.

Angel, J., Maldonado-Sifuentes, C. E., Gelbukh, A. y Sidorov, G. (2021). Developing a language-learning resource for endangered indigenous languages of Mexico. En M. Á. García Trillo, M. L. Sáenz Gallegos, A. G. López Maldonado y A. A. Hurtado Olivares (Coords.), *Procesamiento del lenguaje natural para las lenguas indígenas* (pp. 66-80). Universidad Michoacana de San Nicolás de Hidalgo.

Arzápalo, R. y Gubler, R. (Eds.). (1997). *Persistencia cultural entre los mayas frente al cambio y la modernidad*. Universidad Autónoma de Yucatán.

Austin, P. K. (2016). Language documentation 20 years on. En L. Filipović y M. Pütz (Eds.), *Endangered languages and languages in danger: Issues of documentation, policy, and language rights* (pp. 147-170). John Benjamins.

Blaha Pfeiler, B. y Hofling, A. (2006). Apuntes sobre la variación dialectal en el maya yucateco. *Península*, 1(1), 27-44.

Briceño Chel, F. (2017). U túumben bejilo'ob maayat'aan: los nuevos caminos de la lengua maya: Entre pérdida y revitalización. *Zeitschrift für romanische Philologie*, 133(4), 998-1013.

Briceño Chel, F. y Can Tec, G. R. (Eds.). (2014). *U un'ukbesajil u ts'íibta'al maayat'aan. Normas de escritura para la lengua maya*. Instituto Nacional de Lenguas Indígenas.

Brody, Michal. (2004). *The fixed word, the moving tongue: Variation in written Yucatec Maya and the meandering evolution toward unified norms* [Tesis de Doctorado, Universidad de Texas en Austin].

Brugman, H. y Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. En M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa y R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)* (pp. 2065-2068). Max Planck Institute for Psycholinguistics.

Can Canul, C. y Gutiérrez Bravo, R. (2016). *Narraciones mayas de Campeche (Maayáaj tsikbalilo'ob Kaampech)*. Instituto Nacional de Lenguas Indígenas.

Crasborn, O. y Sloetjes, H. (2014). Improving the exploitation of linguistic annotations in ELAN. En N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk y S. Piperidis (Eds.), *Proceedings of LREC 2014: 9th International Conference on Language Resources and Evaluation* (pp. 3604-3608).

Crystal, D. (1991). *The Cambridge encyclopedia of language*. Cambridge University Press.

Guerrettaz, A. M., Johnson, E. J. y Ernst-Slavit, G. (2020). La planificación lingüística del maya yucateco y la educación bilingüe en Yucatán. *Education Policy Analysis Archives*, 28(132-134), 1-24.

Hernández Méndez, E. (2019). La variación en maya yucateco: un estudio descriptivo desde la dialectología perceptual. *Estudios de Lingüística Aplicada* 69: 143-165.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? En J. Gippert, N. P. Himmelmann y U. Mosel (Eds.), *Essentials of language documentation* (pp. 1-30). De Gruyter.

INEGI. (2020). *Lenguas indígenas y hablantes de 3 años y más, 2020*. http://cuentame.inegi.org.mx/hipertexto/todas_lenguas.htm

Jurafsky, D. y Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Leech, G. (1991). The state of the art in corpus linguistics. En K. Aijmer y B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honor of Jan Svartvik* (pp. 8-29). Longman.

Lehmann, C. (2018). Variación y normalización de la lengua maya. *Cuadernos de Lingüística de El Colegio de México*, 5(1), 331-387.

León-Portilla, M. (2004). El destino de las lenguas indígenas de México. En I. Guzmán Betancourt, P. Maynes y M. León-Portilla (Eds.), *De historiografía lingüística e historia de las lenguas* (pp. 51-70). Siglo XXI Editores.

Mager, M., Gutierrez-Vasques, X., Sierra, G. y Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. En *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 55-69). Association for Computational Linguistics.

Manning, C., Raghavan, P. y Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1), 100-103.

Nava L., E. F. (2021). Reflexiones y razones sobre la catalogación de las lenguas indígenas mexicanas. *Káñina*, 45(1), 109-140.

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada*, 46(1), 93-119.

Pellicer, D. (1999). Derechos lingüísticos y supervivencia de las lenguas indígenas. En A. Herzfeld y Y. Lastra (Eds.), *Las causas sociales de la desaparición y mantenimiento de las lenguas en las naciones de América* (pp. 1-19). Universidad de Sonora.

Pfeiler, B. (1997). El maya: una cuestión de persistencia o pérdida cultural. En R. Arzápalo Marín y R. Gubler (Eds.), *Persistencia cultural entre los mayas frente al cambio y la modernidad* (pp. 55-77). Universidad Autónoma de Yucatán.

Pfeiler, B. (1998). El xe'ek' y la hach maya: cambio y futuro del maya ante la modernidad cultural en Yucatán. En A. Koechert y T. Stolz (Eds.), *Convergencia e individualidad: Las lenguas mayas entre hispanización e indigenismo* (pp. 125-140). Verlag für Ethnologie.

Reyes Pérez, A. y García Zúñiga, H. A. (2021). Hacia el desarrollo de un corpus oral en lengua amuzga. En M. Á. García Trillo, M. L. Sáenz Gallegos, A. G. López Maldonado y A. A. Hurtado Olivares (Coords.), *Procesamiento del lenguaje natural para las lenguas indígenas* (pp. 132-144). Universidad Michoacana de San Nicolás de Hidalgo.

Rojas Curieux, T. (Comp.). (2017). *Corpus lingüísticos: Estudio y aplicación en revitalización de lenguas indígenas*. Universidad del Cauca.

Sánchez Arroba, M. E. (2009). Migración y pérdida de la lengua maya en Quintana Roo. En M. S. Vargas Paredes (Ed.), *Migración y políticas públicas* (pp. 397-468). Universidad de Quintana Roo.

Sierra Martínez, G. E. (2017). *Introducción a los corpus lingüísticos*. Universidad Nacional Autónoma de México.

Sierra, G., Bel Enguix, G. y Mota Montoya, M. (2021). El corpus paralelo de lenguas mexicanas (CPLM). En M. Á. García Trillo, M. L. Sáenz Gallegos, A. G. López Maldonado y A. A. Hurtado Olivares (Coords.), *Procesamiento del lenguaje natural para las lenguas indígenas* (pp. 112-131). Universidad Michoacana de San Nicolás de Hidalgo.

Sima Lozano, E. G., Perales Escudero, M. D. y Be Ramírez, P. A. (2014). Actitudes de yucatecos bilingües de maya y español hacia la lengua maya y sus hablantes en Mérida, Yucatán. *Estudios de Cultura Maya*, 43(8), 157-179.

Sobrino Gómez, M. y Paz Ávila, L. (2008). La condición actual de la lengua maya en Yucatán. *Archipiélago*, 16(60), 41-42.

UNESCO. (2003). *Vitalidad y peligro de desaparición de las lenguas*. http://www.unesco.org/new/fileadmin/multimedia/hq/clt/pdf/Ive_Spanish_edited%20for%20publication.pdf

Valiñas Coalla, L. (2020). *Lenguas originarias y pueblos indígenas de México: Familias y lenguas aisladas*. Academia Mexicana de la Lengua.

Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.

Villanueva, N. (2008). La revaloración de la cultura maya en Yucatán. *Temas Antropológicos*, 30(2), 79-108.

Vinogradov, I. (2016). Linguistic corpora of understudied languages: Do they make sense? *Káñina*, 40(1), 116-130.