

**Cómo citar:** Birchall, J. (2023). Lexicografía y elaboración de diccionarios. En P. Alandia Mercado (Ed.), *Introducción a la Lingüística: Curso para investigadores de lenguas indígenas de Bolivia* (1ª ed., pp. 141-159), Página y Signos/Funprociib Andes. <https://doi.org/10.5281/zenodo.11111121>

## CAPÍTULO 6

# LEXICOGRAFÍA Y ELABORACIÓN DE DICCIONARIOS

Joshua Birchall\*

### 0. Introducción<sup>1</sup>

Los diccionarios son uno de los principales componentes en la descripción de una lengua, junto con las colecciones textuales y las gramáticas descriptivas. La elaboración de un diccionario es una tarea monumental que se basa en la documentación del conocimiento léxico de los hablantes nativos de la lengua. Un diccionario puede cumplir una gran variedad de funciones, como herramienta en la educación, base de estudios científicos del léxico de una lengua o como referencia general para la comunidad lingüística. Un diccionario puede tener un alcance restringido y centrarse en una sola área de conocimiento léxico, tal como *La Ventosa Diidxazá Lexico-Botanical Dictionary* (Pérez Báez y Kaufman, 2019), o puede tener un alcance general y tratar de proporcionar una amplia gama de información lingüística en diferentes dominios semánticos, como el primer diccionario de la variedad cusqueña del quechua de González Holguín (1608).

Los diccionarios han existido durante muchos milenios siendo los primeros ejemplos conocidos los diccionarios bilingües de la lengua sumeria durante el Imperio acadio en Asia occidental, alrededor del 2600 a. C. (van Sterkenburg, 2003). Sin embargo, no fue hasta el siglo XX, con la publicación del *Manual de Lexicografía*

\* Es profesor en la Universidad de Nuevo México. Tiene un doctorado en lingüística por la Universidad Radboud Nijmegen. Sus intereses se centran en la documentación, la descripción y la comparación de lenguas indígenas de Sud América. Está involucrado en una serie de iniciativas para desarrollar diccionarios multimedia y una enciclopedia cultural con comunidades que hablan lenguas en peligro de extinción en el suroeste de la Amazonía. Otras investigaciones cuyas en curso se relacionan con la lingüística histórica, como los métodos filogenéticos, la reconstrucción comparativa de la gramática y el léxico y el desarrollo de bases de datos léxicas.

1 Agradezco a Ester Torrico por la ayuda en traducir una versión preliminar de este artículo del inglés. También agradezco a Gladys Camacho y a los participantes del I Congreso Nacional de Lexicología y Lexicografía de Bolivia por sus comentarios.

de Zgusta (1971) que la lexicografía pasó a ser considerada generalmente como una subdisciplina dentro de la lingüística. En este sentido, Zgusta define el diccionario:

Un diccionario es una lista ordenada sistemáticamente de formas lingüísticas socializadas y compiladas a partir de los hábitos de habla de una comunidad de habla determinada y comentada por el autor de tal manera que el lector cualificado comprenda el significado... de cada forma por separado y es informado de los hechos pertinentes relativos a la función de esa forma en su comunidad. (Zgusta, 1971, p. 17)

Esta definición deja en claro que un diccionario es mucho más que una simple lista de palabras, es una recolección sistemática de formas lingüísticas combinada con información sobre los usos sociales, semánticos y gramaticales de estas formas. Una serie de informaciones y convenciones se han convertido en práctica habitual entre los lexicógrafos, y los recientes avances en la forma en que se llevan a cabo los estudios lingüísticos, así como el mayor acceso a la tecnología digital, han dado forma a la lexicografía en las últimas décadas.

Este capítulo tiene como objetivo proporcionar una descripción general de cómo se puede llevar a cabo un proyecto de diccionario para que sirva de guía práctica sobre cómo reunir y organizar los datos, y cómo estructurar una entrada léxica y el diccionario en general. La discusión aquí se centra en el diseño de diccionarios para comunidades lingüísticas en peligro de extinción, ya que esta es la situación actual de la mayoría de las lenguas indígenas de Bolivia. Los ejemplos se basan en la experiencia del autor al realizar un proyecto de diccionario multimedia en la lengua moré-kuyubim, en peligro de extinción, hablada en el departamento del Beni por la comunidad Moré, y también en otros trabajos lexicográficos en Bolivia.

## 6.1. Preparación para un proyecto de diccionario

Para comenzar un proyecto de diccionario, hay dos preguntas importantes que afectan al diseño de todo el proyecto: ¿quién utilizará el diccionario y cómo lo utilizará? Las respuestas a ambas preguntas afectan el diseño general del diccionario.

### 6.1.1 ¿Quién utilizará el diccionario?

Una de las primeras cosas que se debe tener en cuenta antes de preparar un diccionario es en qué lengua se va a escribir. Si la comunidad de la lengua meta tiene una tradición bien arraigada de escribir en su lengua y si los usuarios previstos del diccionario son hablantes nativos de esta, es habitual elaborar un **diccionario monolingüe** en dicha lengua. En este caso, la lengua que se describe, lo que podemos llamar **lengua meta**, es la misma que se utiliza para describir las formas lingüísticas, lo que podemos llamar **metalenguaje**. Los diccionarios monolingües

son especialmente comunes en los programas de educación nacional, pero no siempre son útiles para el usuario que aún intenta aprender la lengua. Por otro lado, si una comunidad todavía utiliza la lengua como medio de comunicación diario y solo tiene acceso a un **diccionario bilingüe**, un diccionario monolingüe puede ser un aporte bienvenido en la recolección de materiales disponibles sobre la lengua y puede ayudar a aumentar su conocimiento y prestigio.

En situaciones en las que los usuarios esperados de un diccionario no son hablantes nativos y utilizan el diccionario principalmente para aprender más sobre el léxico de la lengua meta, puede ser apropiado utilizar la variante estándar de comunicación diaria en la comunidad como el metalenguaje. En Bolivia, especialmente en las tierras bajas, muchas comunidades indígenas suelen utilizar una variedad local del castellano para la comunicación diaria y especialmente para la comunicación escrita, siendo la lengua indígena la más utilizada de forma hablada y/o por la generación de adultos mayores de la comunidad. Por ejemplo, en la comunidad Moré del departamento de Beni, esta generación es la que utiliza activamente la lengua indígena mientras que la nueva generación tiene un conocimiento pasivo de la lengua porque se comunica principalmente en castellano. Además, existe una variedad de la lengua que se habla tradicionalmente al otro lado del río Itenez/Guaporé por el grupo indígena Kuyubim. Para ayudarlos en el proceso de recuperación y revitalización de su lengua dentro de la comunidad, el diccionario de la lengua moré-kuyubim incluye tanto el castellano y el portugués como los metalenguajes.

Como se señala en Crevels (2012, p. 173), los datos del censo entre los grupos indígenas en Bolivia (y en otros lugares) a menudo no distinguen entre las personas que realmente utilizan la lengua en su vida cotidiana y las que desean aprenderla o se consideran a sí mismas como hablantes de la lengua sobre la base a la identidad étnica. Si el compilador del diccionario fuese o no miembro de la comunidad de habla, esto pone de manifiesto la necesidad de que el trabajo del diccionario se lleve a cabo con una estrecha colaboración de la propia comunidad para comprender mejor y atender a su situación sociolingüística específica y a sus necesidades educativas. Es importante señalar que los diccionarios pueden ser de utilidad no solo para los hablantes de la lengua y los activistas lingüísticos, sino que también pueden ser útiles para los educadores como base de materiales pedagógicos o para los miembros de la comunidad que deseen materiales de referencia sobre su lengua de herencia (Kroskrity, 2015, p. 140). El público al que va dirigido el diccionario y su uso previsto también afectan a todas las etapas del proceso de elaboración del diccionario, desde la decisión de qué documentar hasta la forma de organizar y presentar las entradas léxicas y el propio diccionario, como se expone en las siguientes secciones.

### 6.1.2. ¿Cómo utilizará la gente el diccionario?

Un componente adicional del proceso de elaboración del diccionario, que debe tenerse en cuenta al momento de planificar el proyecto del diccionario, es la forma en la que se utilizará y distribuirá principalmente entre la comunidad de usuarios. Los medios impresos tienen costos adicionales de producción y distribución, pero muchas comunidades de hablantes suelen atribuirles un valor adicional a los materiales impresos en comparación con los digitales. Los materiales impresos pueden distribuirse fácilmente en copias digitales y en forma de documentos electrónicos, como archivos PDF, lo que constituye una forma especialmente útil de recibir comentarios sobre la versión preliminar del diccionario. Sin embargo, una vez que se produce una versión final en forma impresa, suele ser difícil hacer revisiones a los materiales debido a los altos costos de reimpresión. Muchos editores comerciales no permiten la distribución gratuita de copias digitales o impresas de la versión final del diccionario una vez publicado, pero actualmente existen iniciativas en la publicación con acceso abierto que permiten la descarga gratuita de copias digitales de los diccionarios y una opción de impresión bajo demanda de pago para quienes prefieran una copia impresa. Las instituciones públicas como los museos y las universidades pueden permitir la publicación de materiales impresos a través de opciones de financiamiento no comercial para su distribución gratuita, especialmente para materiales destinados a la comunidad de usuarios, pero generalmente con un número más limitado de ejemplares.

Los diccionarios completamente digitales se están convirtiendo en una opción más frecuente entre los lexicógrafos y los usuarios de diccionarios. Especialmente en los casos en los que no hay una gran comunidad de hablantes que utilice la lengua a diario, la capacidad de incorporar la pronunciación de los lexemas, así como los diferentes ejemplos del uso de estos lexemas como los archivos multimedia pueden ser indispensables para los estudiantes de lenguas. Un beneficio adicional de los diccionarios digitales, a diferencia de los medios impresos, es que es fácil producir versiones revisadas, ya sea para ampliar el material original publicado o para revisar los registros. Sin embargo, es posible que no todos los miembros de la comunidad tengan acceso a la tecnología adecuada para utilizar el diccionario con regularidad. Los materiales con alto contenido multimedia almacenados en un sitio web o una aplicación a menudo requieren conexiones a internet rápidas y estables para su uso óptimo. Ya sea que estén establecidos en la web o en aplicaciones, los recursos completamente digitales tienen algunos inconvenientes adicionales, como la necesidad de actualizar la versión publicada para garantizar la compatibilidad con los estándares tecnológicos cambiantes y los diferentes sistemas operativos de las computadoras personales y de los dispositivos móviles. Mientras que un libro impreso puede seguir cumpliendo su función durante décadas, sin una actualización

periódica, un producto digital puede volverse funcionalmente obsoleto. Los lexicógrafos pueden explorar diferentes opciones de publicación que mejor se adapten a las necesidades de los usuarios. Como se expone en la siguiente sección, una metodología basada en la elaboración de un diccionario a partir de una base de datos léxica bien diseñada permite una mayor flexibilidad en las opciones de publicación y garantiza una mayor longevidad y utilidad de los datos lingüísticos que constituyen la base del propio diccionario.

## 6.2. Documentación léxica

Al comenzar la fase de documentación léxica de un proyecto de diccionario, se debe tener en cuenta una serie de preguntas: ¿qué materiales ya están disponibles?, ¿cómo se pueden recopilar los datos primarios? y ¿cómo organizar los datos una vez recopilados? La comprensión clara de estas preguntas permite a un lexicógrafo planificar y ejecutar adecuadamente la fase de documentación léxica de un proyecto de diccionario.

### 6.2.1. ¿Qué materiales ya están disponibles?

Cuando se planea desarrollar la documentación del léxico de una lengua, es importante estudiar los materiales existentes de esta. Dicha documentación existente puede abarcar desde notas de campo escritas a mano hasta gramáticas descriptivas publicadas y colecciones de documentación multimedia de acceso público, almacenadas en archivos profesionales como el Archivo de Lenguas en Peligro de Extinción (ELAR) o el Archivo de Lenguas Indígenas de América Latina (AILLA). Tener un buen conocimiento de los materiales disponibles puede ayudar a que la fase de documentación léxica de un proyecto de diccionario sea más eficiente al proporcionar una línea de base que ayuda a orientar las prioridades de documentación.

Una vez conocida la documentación disponible, la lexicografía debe considerar cómo se puede utilizar la documentación anterior para mejorar la calidad y la cobertura del propio diccionario, así como para facilitar el proceso de documentación léxica. En algunos casos, como en los materiales audiovisuales existentes, puede ser posible incorporar directamente al diccionario en desarrollo partes del trabajo de documentación anterior, dependiendo de los derechos intelectuales adscritos a estos materiales y de sus condiciones de uso, y, como siempre, citando y atribuyendo la autoría adecuadamente. En otro caso, puede ser más útil usar los materiales anteriores para orientar las nuevas sesiones de documentación. A continuación, se examinarán dos ejemplos diferentes de cómo los materiales lingüísticos existentes pueden guiar el proceso de documentación léxica.

Por ejemplo, cuando el autor comenzó el trabajo de campo con los Moré en 2017 para elaborar un diccionario multimedia, estaba familiarizado con aspectos generales de la cultura material y de la estructura social de los grupos vecinos de

Chapacura en Brasil, como los Wari' y los Oro Win, pero tenía poco conocimiento de los elementos culturales específicos de los Moré. Sin embargo, a través de una consulta cuidadosa a una serie de trabajos etnográficos, como en Ryden (1942) y Leigue Castedo (1957), y gracias a que este último incluye también una modesta lista de palabras, fue posible identificar aspectos específicos de la cultura y de la historia de los Moré que, aunque no se ven ni se discuten con frecuencia en el contexto moderno de la sociedad Moré, son importantes para el pueblo, para su historia y su forma de vida tradicional. Después de discutir sobre los materiales con varios colaboradores de Moré que estaban vivos durante el período de contacto con la sociedad boliviana en general, a principios y mediados del siglo XX, fue posible realizar entrevistas de elicitación para registrar estas palabras y los ejemplos de su uso (véase Birchall et al., 2022).

Otro ejemplo reciente del uso de materiales existentes para la lexicografía moderna puede verse en Danielsen et al. (2019) con la lengua guarayo. El equipo de investigación adaptó el diccionario publicado en Hoeller (1932) a la ortografía comunitaria actual y luego trabajó con hablantes jóvenes de la lengua para corregir cualquier error y grabar el audio de las diferentes entradas. El proyecto se publicó con notas históricas, culturales y lingüísticas adicionales en la revista en línea de acceso abierto *Dictionaria*.

### 6.2.2. ¿Cómo se puede recolectar los datos primarios?

Como se señala en Frawley et al. (2002, p. 1), el proceso de documentación léxica y la elaboración de diccionarios avanza “expandiéndose desde una modesta lista de palabras y glosas a algo parecido a una enciclopedia cultural”. Incluso el estudio más preliminar de una lengua suele implicar la recolección de elementos de vocabulario básicos, y es normalmente ahí dónde y cómo comienza el proceso de documentación léxica. A medida que el proceso de documentación avanza, es esencial que el lexicógrafo atienda a la comunidad de futuros usuarios construyendo una colección de entradas léxicas que refleje sus necesidades comunicativas. Para ello, es necesario comprender no solo el contexto físico en el que se utiliza la lengua, como la flora, la fauna y los fenómenos naturales típicos relevantes para la vida cotidiana de la comunidad, sino también el contexto social en el que se habla la lengua. Especialmente cuando el lexicógrafo no es miembro de la comunidad de habla, el conocimiento cultural e histórico de fondo es necesario para alcanzar una profunda y adecuada cobertura léxica que debe adquirirse a través de la observación participante, las entrevistas con miembros de la comunidad bien informados y el estudio cuidadoso del registro etnolingüístico existente.

La documentación del léxico puede utilizar diversas metodologías. Cuando el proyecto del diccionario se lleva a cabo junto con un proyecto de documentación

lingüística más amplio, o como una fase posterior de este, es posible que una parte sustancial del conocimiento léxico ya se haya documentado en un corpus transcrito de habla natural. Los casos de uso del lenguaje dentro de un corpus documental multimedia pueden utilizarse como ejemplos naturalistas del uso de elementos léxicos específicos. Sin embargo, la elaboración de un corpus de este tipo suele ser un proceso largo y prolongado que puede durar muchos años y, por lo tanto, puede no ser factible para proyectos cuyo objetivo principal sea la lexicografía. A menudo, es necesario emplear otros métodos para garantizar una cobertura adecuada del léxico y, a menudo, se utilizan diversas variaciones de las técnicas estándar de **elicitación**.

La **elicitación lingüística** es “la recolección de respuestas a estímulos lingüísticos o no lingüísticos diseñados para estudiar la competencia lingüística de los encuestados y/o sus prácticas de uso de la lengua” (Bohnenmeyer, 2014, p. 21). La práctica prototípica de elicitación es aquella en la que un investigador le pide a un hablante que proporcione una traducción cercana de una palabra o frase específica en la lengua meta que ha sido proporcionada por el investigador en una metalengua. Para fines lexicográficos, las entrevistas de elicitación organizadas en torno a dominios semánticos específicos a menudo producen los mejores resultados, en lugar de tratar de obtener palabras por orden alfabético o por parte del discurso. Si bien a menudo es una sociedad fructífera, especialmente para la documentación de lexemas específicos de forma aislada, basarse únicamente en los equivalentes de traducción puede limitar el alcance de la cobertura léxica, ya que el lexicógrafo suele limitarse a los conceptos presentes en el metalenguaje. También se debe tomar en cuenta los efectos que la estructura del metalenguaje ejerce sobre la lengua meta a través de la **imposición estructural**, el **calco semántico** y otros efectos de **contacto**, especialmente en la producción de enunciados de más de una sola palabra (véase en Lüpke, 2009 y Payne, 1997, pp. 366-371 para debates útiles sobre el papel de la elicitación y la colección de textos en el trabajo de campo lingüístico). La definición anterior de elicitación lingüística también incluye el método cada vez más común de utilizar estímulos visuales como imágenes o vídeos para obtener una respuesta lingüística de un hablante de la lengua. Como señala Majid (2012), el uso de estímulos visuales ayuda a limitar la influencia del metalenguaje en la respuesta lingüística obtenida. Una técnica adicional que es, en cierto modo, una hibridación, tanto de la elicitación por equivalente de traducción como de la elicitación por estímulos visuales, es la práctica de llevar a cabo la documentación lingüística a través de lo que Lüpke (2009) denomina “eventos comunicativos escenificados”, donde se pide a los hablantes que manifiesten su comportamiento lingüístico esperado como participantes en una situación específica. Todas estas diferentes técnicas pueden contribuir a enriquecer la documentación léxica disponible de una lengua y pueden complementar los conocimientos lingüísticos adquiridos mediante la colección de textos y la observación de los participantes.

Por razones prácticas, la documentación de una lengua que no ha sido bien documentada o descrita comienza inicialmente con una mayor dependencia de las técnicas tradicionales de elicitación y luego pasa a la recopilación de textos y a los métodos de observación participante, especialmente cuando el investigador no es miembro de la comunidad de habla. A medida que el investigador construye gradualmente un corpus de conocimiento léxico, es esencial que se sigan protocolos estándar en la documentación lingüística, como la recopilación oportuna de metadatos (o, mejor dicho, información sobre cómo se recolectó esa misma información lingüística), así como la transcripción de los datos a una forma ortográfica que permita una fácil búsqueda de los contenidos. Los metadatos a menudo se catalogan como archivos de texto u otros archivos de formato digital que acompañan a los archivos multimedia dentro de un corpus de documentación. Las transcripciones se suelen realizar con programas informáticos que permiten la alineación temporal entre los medios y el texto, como ELAN o SayMore. Una vez recogidos los datos y sometidos al tratamiento necesario para garantizar su utilidad, estos pueden organizarse en una base de datos léxica que puede utilizarse específicamente para la elaboración de un diccionario.

### **6.2.3. ¿Cómo organizar los datos una vez recolectados?**

A medida que un lexicógrafo recopila un corpus cada vez más amplio de información léxica y ejemplos de uso, se recomienda organizar estos datos en algún tipo de base de datos. En los últimos años se han introducido una serie de buenas prácticas en el desarrollo de bases de datos (ver Capítulo 1) procedentes de disciplinas ajenas a la lingüística, como el desarrollo de software, las estadísticas y la administración de empresas, que pueden aplicarse fácilmente a las bases de datos lingüísticos. En este sentido, Wickham (2014) resume varias de estas prácticas bajo la denominación de principios de “datos ordenados”, es decir: (i) cada variable forma una columna, (ii) cada observación forma una fila y (iii) cada tipo de unidad de observación forma una tabla. En su forma más simple, podemos imaginar una base de datos léxica como una hoja de cálculo y podemos interpretar estos principios en el sentido de que cada fila en la tabla de la hoja de cálculo corresponde a un lexema concreto (una “observación”) y que cada columna de la tabla corresponde a una propiedad específica de ese lexema (una “variable”), como parte del discurso, la clase semántica, la transcripción fonética, la transcripción ortográfica, etc. (véase la sección 6.3., más adelante, para examinar la información que suele asociarse a una entrada léxica). Es importante que cada observación de una tabla tenga su propio número de identificación para que pueda ser referenciada en otras secciones de la base de datos. Para incluir ejemplos de uso vinculados a lexemas específicos, es posible hacer referencia a su número de identificación en una tabla separada donde se recopilan y seleccionan los datos textuales, o bien reproducir los propios ejemplos

dentro de la tabla de lexemas. Cuando se elabora un diccionario digital en lugar de uno impreso, casi siempre se utiliza una base de datos para elaborar el producto final. Bergenholtz y Nielsen (2013) ofrecen comentarios adicionales sobre el diseño de bases de datos lexicográficas que pueden dar cuenta de relaciones más complejas entre formas y sus significados, especialmente los casos de sinonimia (múltiples formas con el mismo significado) y polisemia (una única forma con múltiples significados).

Existen diversos programas para lingüistas de campo que permiten la elaboración de bases de datos de información textual y léxica. Un método común es el uso de software de trabajo de campo producido por el Instituto Lingüístico de Verano (ILV/SIL), originalmente Shoebox, que luego fue reemplazado por Toolbox, y ambos han sido sustituidos por Fieldworks Language Explorer, comúnmente conocido como FLE<sub>x</sub>. Estos programas cuentan con una serie de herramientas que permiten la recopilación de bases de datos de información léxica derivada directamente de los datos textuales y, además, cuentan con una serie de funciones que proporcionan glosas semiautomáticas de textos recién agregados en función de la información suministrada con anterioridad. Especialmente en el caso de FLE<sub>x</sub>, algunos investigadores han tenido dificultades para exportar datos léxicos del programa de forma que puedan utilizarse fácilmente para otras aplicaciones.

A la hora de diseñar una base de datos léxica, es importante sopesar las ventajas adicionales de utilizar un software de terceros frente a las desventajas de no disponer de los datos en un formato textual fácilmente transportable y abierto, como los valores separados por comas (.csv) o los valores separados por tabulaciones (.tsv) que se utilizan habitualmente en las disciplinas digitales, aunque algunos software de terceros permiten exportar los datos en dicho formato. Otra opción para almacenar sus datos léxicos es utilizar el Cross-linguistic Data Formats (Forkel et al., 2018), un formato para guardar diferentes tipos de datos lingüísticos diseñados para garantizar la compatibilidad entre herramientas, plataformas y programas existentes. El CLDF tiene una serie de principios de diseño que respaldan esta multifuncionalidad: todos los datos deben almacenarse como archivos de texto con codificación UTF-8 (unicode), los conjuntos de datos deben ser editables a mano en un software de hoja de cálculo comúnmente disponible y las entidades, a las que se hace referencia, dentro de un archivo de datos deben estar vinculadas a catálogos de referencia como Glottolog (Hammarström et al., 2021) para lenguas/dialectos o como Concepticon (List et al., 2021) para conceptos semánticos. Este recurso requiere de una conservación manual adicional de los datos en comparación con el software disponible en el mercado, pero tiene la ventaja adicional de poder modificar fácilmente los datos y utilizarlos para otros fines.

Pasar directamente de los datos primarios recogidos a un formato de diccionario publicable sin una base de datos intermedia puede parecer un paso que ahorra tiempo

en un principio, pero el uso de una base de datos tiene una serie de ventajas. Una de las principales ventajas de un enfoque de base de datos para la lexicografía es una cuestión de revisión y organización. Disponer de todas las entradas léxicas en el mismo formato de búsqueda permite identificar fácilmente las entradas reduplicadas o, en los casos en los que no existe un consenso comunitario generalizado sobre las convenciones ortográficas, las múltiples instancias del mismo lexema transcritas de diferentes maneras y/o con definiciones o traducciones similares, pero no idénticas. La aplicación de diferentes filtros a los datos tabulares en el software de hoja de cálculo habituales, permite una rápida inspección visual de los datos. Una ventaja adicional de un enfoque de base de datos para la lexicografía es la posibilidad de realizar modificaciones globales en la propia base de datos o en cualquier producto derivado de la misma. Independientemente de la estructura de la base de datos utilizada, es de suma importancia que la información almacenada en ella se conserve en un formato de archivo que garantice la longevidad de los datos.

Un último aspecto sobre la organización de datos primarios para un proyecto de lexicografía es qué hacer con los archivos multimedia que acompañarán a un diccionario digital. Lo más importante es que los archivos multimedia, ya sean de audio o video, tengan una convención de nomenclatura uniforme que se les aplique. Una convención de nomenclatura debe permitir la identificación inmediata del contenido de los archivos, siendo cada nombre de archivo único e independiente de la estructura de carpetas más amplia en la que se almacenan los archivos. Tener un nombre único para cada archivo permite al lexicógrafo referirse a cada archivo específico dentro de la base de datos lexicográfica más amplia. En los casos en los que un archivo multimedia se extrae de un archivo multimedia mayor, como una ficha única de una entrevista de elicitación más amplia, se puede incluir información en el nombre del archivo extraído que permita identificar el archivo de origen o mantener una tabla adicional con los metadatos de cada archivo extraído que incluya su procedencia.

### **6.3. Estructura de una entrada léxica**

Una vez que se ha documentado la información léxica que se incluirá en un diccionario y se han organizado los datos, la siguiente tarea es organizar las entradas reales del diccionario. Esto lleva a la pregunta obvia: ¿qué incluir en una entrada de diccionario?

El contenido de una entrada de un diccionario puede variar en función de los objetivos generales del proyecto y del público al que va dirigido. Garrett (2018, pp. 197-198), cuando menciona los objetivos de un diccionario para apoyar los esfuerzos de revitalización de la lengua, observa que “un diccionario tiene que ayudar a la gente a usar la lengua en peligro de extinción (EL)”. Por ejemplo, cuando varias

palabras de la lengua en peligro de extinción (EL) tienen significados similares o traducciones idénticas en la lengua materna, un diccionario para la revitalización de la lengua debe ayudar a los usuarios a comprender las implicaciones de sus elecciones de palabras. ¿Qué palabra es apropiada en distintos contextos de uso? ¿En qué se diferencian desde el punto de vista semántico, pragmático y sintáctico? Los estudiantes que se basan en un diccionario que carece de detalles sobre estos parámetros de uso pueden tender a combinar el vocabulario de la lengua en peligro de extinción (EL) con patrones sintácticos y semánticos de la lengua materna, dando lugar a una lengua materna “relexificada”.

Aunque un lexicógrafo puede tomar diferentes decisiones sobre cómo organizar y presentar los datos lingüísticos dependiendo de si los usuarios previstos normalmente utilizarán el diccionario para el aprendizaje de idiomas en lugar de material de referencia general, existen componentes que son estándar en la lexicografía moderna que se encontrarán en la mayoría de las entradas del diccionario, como el **lema** identificado, una **pronunciación**, una **definición**, la **parte de la oración**, **información gramatical** y **ejemplos de uso**. Muchos diccionarios también optan por incluir **información cultural o etimológica** adicional cuando está disponible.

El **lema** de una entrada de diccionario, a menudo llamado también **encabezado o término**, es la forma canónica del lexema a la que se asignarán todas las demás variantes morfológicas de la palabra. Otra forma de describir un lema es como “una abstracción que subsume otras formas relacionadas tanto morfológica como semánticamente” (Frawley et al., 2002, p. 3). Por ejemplo, en la entrada de la palabra en inglés *break*, se debe incluir también la forma de pasado *broke* y la forma del participio *broken*. Bajo esta misma entrada cubierta por el lema *break*, también se pueden incluir formas relacionadas adicionales, como desglosar y romper. Estas formas relacionadas ayudan a los usuarios del diccionario a encontrar fácilmente la información que buscan.

La estructura morfológica de la lengua que se describe y las convenciones comúnmente usadas por la comunidad de hablantes deberían ayudar a determinar la selección de un lema para una entrada. Independientemente de las elecciones que se realicen, que se puede considerar la “ideología del lenguaje” del diccionario, lo más importante para los usuarios del diccionario es que estas opciones se apliquen de forma coherente en todas las entradas, y que estas opciones se muestren explícitamente en la portada del diccionario (Rhodes et al., 2018). Es habitual que las diferentes formas flexionadas de un mismo lexema se organicen bajo el mismo lema, mientras que los lexicógrafos varían en cuanto a si las formas derivadas deben organizarse bajo el mismo lema. Otro tema por considerar es si el propio lema debe llevar alguna inflexión en su forma presentada al inicio de una entrada. Por un lado, una **inflexión** en la forma del prefijo haría que todas las entradas de la misma parte de la oración comenzaran con la misma letra. Por otro lado, la ausencia total de inflexión en un

lema podría dar lugar a una forma que no se da de manera natural en la lengua, lo que dificultaría que el usuario localice la información que busca.

Debajo del encabezado de un lema, la entrada debe incluir su pronunciación e información gramatical adicional, como parte de la oración. La pronunciación se presenta generalmente como una aproximación fonética estandarizada utilizando el Alfabeto Fonético Internacional. Algunas tradiciones lexicográficas de lenguas sin ortografía fonémica como el inglés pueden adoptar convenciones diferentes, y las lenguas con una ortografía fonémica transparente, pero con una considerable variación dialectal en la pronunciación, pueden optar por no incluir ninguna pronunciación, como en el *Diccionario de la lengua española* (Real Academia Española, 2001). Las categorías gramaticales específicas de la lengua suelen incluirse en una entrada del diccionario. Las categorías generales de partes de la oración, como sustantivo, verbo o adjetivo, son las más comunes, pero también se incluyen categorías específicas de la lengua, como gerundios o pronombres interrogativos, así como otra información gramatical relevante para la entrada concreta de la lengua, como la clase/género de los sustantivos o la clase flexiva de los verbos. Algunos diccionarios también incluyen los paradigmas de las formas flexionadas, especialmente cuando existen patrones múltiples o irregulares en la lengua.

Las entradas del diccionario siempre incluyen una definición del lema y sus significados asociados. Es una práctica habitual definir una palabra sin utilizar la propia palabra para evitar la redundancia. Como señala Geeraerts (2003), una palabra puede tener múltiples significados o sentidos más allá de su denotación, es decir, el tipo de elementos en el entorno a los que se refiere la palabra, como su significado emotivo, su significado gramatical y su significado pragmático. Estos diferentes significados de una palabra se basan en gran medida en el uso que hacen los hablantes de la misma y es la labor del lexicógrafo describir estos diferentes usos. En los diccionarios bilingües, a veces se recurre a la traducción directa cuando existen conceptos correspondientes entre la lengua meta y el metalenguaje. Sin embargo, puede ser problemático suponer que, incluso en los casos en que existe una traducción directa entre las dos lenguas, el uso de los términos es idéntico y el lexicógrafo debe presentar los diferentes sentidos en los que se utiliza la palabra. Tanto en los diccionarios monolingües como en los bilingües, también se pueden incluir palabras relacionadas, como **sinónimos** y **antónimos**, o **expresiones idiomáticas** específicas de la cultura que hacen uso de la palabra.

Varios diccionarios incluirán ejemplos de uso del lema descrito en sus diferentes acepciones. Como se destaca en la cita de Garrett (2018) al comienzo de esta sección, el objetivo de los ejemplos es ayudar a guiar al usuario del diccionario en el uso del término. Puede ser útil para el usuario incluir ejemplos que resalten otras palabras asociadas con la entrada. Por ejemplo, en una entrada del diccionario

sobre la palabra ‘miel’, sería útil que el usuario tuviera un ejemplo que incluyera el verbo apropiado para el consumo de miel, como por ejemplo si la miel se “come”, se “toma”, se “lame”, se “bebe” o de otra manera (todos estos son verbos correspondientes al consumo de miel que el autor ha encontrado en su trabajo de campo con diferentes lenguas en la Amazonía). Además de los ejemplos de uso de un elemento léxico, muchos lexicógrafos incluyen imágenes del referente cuando es posible. Si el diccionario es multimedia, también se pueden incluir videos o audios que proporcionen un ejemplo del uso del término o un ejemplo del uso del referente del término en su contexto apropiado.

Un último dato que suele incluirse en una entrada de diccionario es lo que se sabe sobre la etimología de un término en particular. En las lenguas con una rica tradición filológica, la historia de una palabra puede ser bien entendida y atestiguada a través de documentos históricos de diferentes épocas, así como en reconstrucciones propuestas para el término en protolenguas anteriores. Sin embargo, en el caso de las lenguas que carecen de tal tradición, es posible incluir simplemente términos de otras lenguas que puedan identificarse como afines.

#### **6.4. Estructura del diccionario**

Ya hemos mencionado sobre la planificación de un proyecto de diccionario, de la recopilación de datos y la organización de esa información en entradas, pero ahora pasamos a dos preguntas sobre la estructura general del diccionario: ¿cómo se debe representar ortográficamente la lengua meta? y ¿cómo puede organizarse el diccionario?

La pregunta sobre qué ortografía utilizar es a menudo objeto de un acalorado debate en el contexto de la lexicografía de las lenguas en peligro de extinción. En comunidades lingüísticas con una larga tradición de convenciones de escritura, la tarea es bastante sencilla siempre que la ortografía represente suficientemente los sonidos contrastantes relevantes de la lengua. Sin embargo, un diccionario es a veces la primera obra escrita importante en algunas lenguas y puede tener un considerable efecto normalizador sobre la forma escrita de la lengua en su uso futuro. Cuando una lengua no tiene una ortografía establecida y ampliamente adoptada, lo más probable es que los usuarios del diccionario ya estén escribiendo en otra lengua. El diseño de la ortografía plantea dos cuestiones: ¿qué símbolos utilizar? y ¿qué símbolos representan qué sonidos en la lengua?

La mayoría de los lingüistas prefieren utilizar símbolos fonéticos cuando trabajan con una lengua no escrita, sobre todo porque estos símbolos tienen la capacidad de representar aspectos de la lengua que no suelen recogerse en las ortografías basadas en caracteres latinos, como la longitud, la aspiración y el tono.

Sin embargo, desde el punto de vista del usuario, el uso de símbolos puramente fonéticos como la ortografía tiene una serie de inconvenientes, como la falta general de familiaridad en la forma de pronunciar algunos de estos símbolos, así como la dificultad general para escribir esta lengua en una computadora o teléfono sin teclado especial. La mayoría de los lexicógrafos adoptan una ortografía práctica que utiliza el mismo conjunto de símbolos empleados en el idioma oficial, y muchos optan por complementarla con una fonética o fonémica más detallada como parte de la entrada estándar (véase § 6.4.).

Suponiendo que se adopte el mismo conjunto de símbolos que se utilizan en el idioma oficial como base para una ortografía práctica utilizada como medio principal para presentar formas escritas de los lexemas, se plantea entonces la cuestión de cómo representar los sonidos de la lengua meta que no están presentes en el idioma oficial. Los lexicógrafos han adoptado una serie de convenciones, como la representación de las oclusiones glotales o las eyectivas con un apóstrofo y el uso de acentos para representar los contrastes tonales. Independientemente de las elecciones que se realicen, es importante que un lexicógrafo, ya sea de la propia comunidad de habla o no, trabaje en estrecha colaboración con otros miembros de la comunidad que tengan un interés personal en la adopción de la ortografía, como educadores y otros líderes de la comunidad. Las implicaciones de cada elección deben presentarse y debatirse con el objetivo de alcanzar un consenso. En Hinton (2014), se discuten una serie de cuestiones relacionadas con el desarrollo de ortografías comunitarias para lenguas no escritas, donde se destaca que las ortografías pueden tener considerables implicaciones políticas, y es una de las implementaciones más obvias de la ideología lingüística dentro de un diccionario. Hinton identifica una serie de factores importantes que deben considerarse cuidadosamente, el más importante es que la ortografía sea considerada aceptable por la comunidad lingüística. Otros factores incluyen la capacidad de representar con precisión todos los sonidos distintivos de la lengua, la facilidad con la que se puede aprender y la sencillez con la que se puede reproducir en los medios.

La última pregunta por considerar es cómo organizar las entradas de un diccionario. El enfoque más común es presentar las entradas en orden alfabético. Sin embargo, como se discute en Frawley et al. (2002, p. 9), se deben tomar ciertas decisiones a la hora de alfabetizar las entradas dependiendo de los símbolos adoptados en la ortografía, como por ejemplo si las entradas que comienzan con el símbolo *p* irían antes o después de las que tienen *p*, o si un dígrafo como *ch* que representa un solo sonido debería ir antes de las entradas que comienzan con *ci* o bien deberían formar su propia sección del diccionario. Otro enfoque consiste en organizar el diccionario temáticamente en torno a diferentes nociones semánticas

o gramaticales, como los animales o los verbos, con las entradas posteriores organizadas alfabéticamente dentro de cada sección. Esta organización puede ser especialmente útil para los usuarios del diccionario que se dedican principalmente al aprendizaje de lenguas, y es más fácil de aplicar cuando cada lema tiende a asociarse con un solo significado.

Una consideración adicional para la organización de las entradas es si estas deben organizarse según la lengua meta o, en el caso de los diccionarios bilingües, según el metalenguaje. En el caso de los diccionarios multimedia, es habitual permitir una opción para ambas organizaciones en función de la configuración seleccionada por el usuario, pero, en el caso de los diccionarios publicados por escrito, esto puede resultar más complicado. Los usuarios de diccionarios que se preocupan principalmente por aprender el idioma pueden encontrar beneficioso buscar palabras utilizando el metalenguaje. Para apoyar a este tipo de usuarios, una opción razonable puede ser tener la estructura principal del diccionario organizada por la lengua meta con un glosario adicional al final, que está organizado en el metalenguaje. Como ocurre con la mayoría de las opciones en lexicografía, son las necesidades de los usuarios las que deben impulsar el diseño y la implementación del diccionario.

## 6.5. Lexicografía en el contexto boliviano

Hay una larga tradición de lexicografía con las lenguas indígenas de Bolivia incluso antes de la fundación del Estado Plurinacional de Bolivia. Por ejemplo, uno de los primeros estudios lexicográficos con una lengua de América del Sur fue realizado por el padre jesuítico italiano Ludovico Bertonio con hablantes de la lengua aimara que vivían al sur del lago Titicaca, a principios del siglo XVII (Bertonio, 1612). La tradición de lexicografía aimara continúa hasta hoy en universidades bolivianas.

En las tierras bajas bolivianas, también hay una larga tradición lexicográfica desde el periodo colonial, como, por ejemplo, el *Arte de la lengua Moxa con su vocabulario y cathecismo*, publicada originalmente por el jesuita Pedro Marban a principios del siglo XVIII (Marban, 1894 [1701]). Según Adelaar (2012), es posible que muchos otros trabajos lexicográficos y gramaticales hayan sido producidos durante el periodo colonial, como las “artes” y “tesoros” de los jesuitas, pero que se han perdido por solo ser publicados como manuscritos o que aún están por descubrir en los archivos de la Iglesia católica en América Latina o Europa.

En el siglo XX se retomó el interés sobre la descripción de las lenguas indígenas de Bolivia y sus vocabularios, primeramente por investigadores relacionados a diferentes institutos religiosos, como el diccionario de Guayaros

por el franciscano Hoeller (1932), y otros del Instituto Lingüístico del Verano, como el trabajo de Van Wynen y Van Wynen (1962) sobre la lengua tacana.

En las últimas décadas, varios investigadores académicos han avanzado en la descripción lexicográfica de las lenguas bolivianas a partir de la producción de diccionarios y vocabularios. Muchos de estos trabajos sirven como ejemplos de la implementación de lexicografía moderna mediante el uso de datos multimedia para producir diccionarios digitales. El diccionario de Guarayu de Danielsen et al. (2019) mencionado arriba es un buen ejemplo de cómo se puede utilizar materiales lexicográficos antiguos para producir diccionarios modernos. Otro elemento importante de lexicografía moderna es que los investigadores indígenas han tomado un papel central en la documentación léxica y la elaboración de diccionarios de sus lenguas. Por ejemplo, el diccionario *Ese Ejja E Esowijo Ewowi Ja'a Pokiajji* fue producido recientemente por el lingüista Alejandro Machuqui Dominguez, también conocido como Bawapoji, miembro de la comunidad Ese Ejja (Machuqui Dominguez, 2021). Otro ejemplo es el diccionario *Moré-Kuyubim* producido por el autor de este texto, miembros de las comunidades Moré y Kuyubim y otros colaboradores lingüistas y antropólogos (Birchall et al., 2022). Estos dos últimos diccionarios también tienen versiones multimedia que están disponibles como aplicativos Android. Así se puede ver que Bolivia es un líder de lexicografía moderna en América Latina, pero, dado el estado de riesgo de sus lenguas indígenas, aún queda mucho trabajo lexicográfico por hacer.

## 6.6. Conclusión

Como se mencionó a lo largo de este capítulo, es importante que se tome una decisión en la elaboración de un diccionario, ya sea en la selección de la forma morfológica de un lema, la ortografía adoptada o en su organización general, y que esta elección esté motivada por las necesidades de los usuarios y que se aplique de forma coherente en todo el trabajo. El lugar apropiado para presentar estas opciones es la introducción del diccionario. Aquí, el lexicógrafo puede proporcionar información adicional a los usuarios para guiarlos en el uso del diccionario, con la presentación de las motivaciones por las que se adoptaron dichas elecciones. En otras palabras, es una oportunidad para que el lexicógrafo presente la ideología lingüística que subyace en la producción del diccionario (véase Kroskrity, 2004 y Rhodes et al., 2018 para discusiones útiles sobre el papel de la ideología lingüística en el desarrollo del diccionario). La introducción también es una oportunidad para proporcionar más información sobre la lengua, como sus sonidos y su gramática, o información sobre la historia y la cultura de las personas que la hablan, así como para reconocer a todos los colaboradores que han contribuido con el diccionario.

Además, en este capítulo, se han abordado una serie de cuestiones relacionadas con la lexicografía, especialmente en lo que respecta a la elaboración de diccionarios para lenguas en peligro de extinción en el contexto boliviano. Existe una necesidad urgente de aumentar el estudio de los léxicos de todas las lenguas, especialmente las de Bolivia. Una de las mejores maneras de hacerlo es mediante la capacitación de los miembros de las comunidades indígenas en los métodos y la teoría necesarios para documentar y describir sus propias lenguas. Un diccionario es una poderosa herramienta para tomar los conocimientos adquiridos a través de la documentación lingüística y la lexicografía y ponerlos en manos de sus usuarios.

## Referencias

- Adelaar, W. F. H. (2012). Historical overview: Descriptive and comparative research on South American Indian languages. En L. Campbell y V. Grondona (Eds.), *The indigenous language of South America: a comprehensive guide* (pp. 1-57). De Gruyter Mouton.
- Bergenholtz, H. & Nielsen, J. S. (2013). What is a lexicographical database? *Lexikos*, 23, 77-87.
- Bertonio, L. (1612). *Vocabulario de la lengua aymara*. Francisco del Canto.
- Birchall, J., Chyc, P., Leigue Sae, J. B., Cujubim, M., Costa, C. N. d. (2022). *Dicionário Multimídia Moré-Kuyubim. Versión 1.0*. Museu do Índio. <http://japiim.museudoindio.gov.br/dic/morekuyubim/>.
- Bohnemeyer, J. (2014). A practical epistemology for semantic elicitation in the field and elsewhere. En R. Bochnak y L. Matthewson (Eds.), *Methodologies in semantic fieldwork* (pp. 13-46). Oxford University Press.
- Crevels, M. (2012). Language endangerment in South America: The clock is ticking. En L. Campbell y V. Grondona (Eds.), *The indigenous language of South America: a comprehensive guide* (pp. 167-233). De Gruyter Mouton.
- Danielsen, S., Sell, L. y Terhart, L. (2019). Guarayu. A revised dictionary by Alfred Hoeller. *Dictionaria*, 7, 1-3590. <https://dictionaria.clld.org/contributions/guarayu>.
- Forkel, R., List, J.M., Greenhill, S., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A. y Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5, 180205 (2018). <https://doi.org/10.1038/sdata.2018.205>

- Frawley, W., Hill, K. C., y Munro, P. (2002). Making a dictionary: Ten issues. En W. Frawley, K. C. Hill y P. Munro (Eds.), *Making dictionaries: preserving indigenous languages of the Americas* (pp. 1–22). University of California Press.
- Garrett, A. (2018). Online dictionaries for language revitalization. En L. Hinton, L. Huss, y G. Roche (Eds.), *The Routledge handbook of language revitalization* (pp. 197–206). Routledge.
- Geeraerts, D. (2003). Meaning and definition. En P. van Sterkenburg (Ed.), *A practical guide to lexicography* (pp. 83–93). John Benjamins Publishing Company.
- Gonzalez Holguín, D. (1608). *Vocabulario dela lengua general de todo el Peru llamada lengua Quichua, o del Inca: corregido y renovado conforme ala propiedad cortesana del Cuzco*. Francisco del Canto.
- Hammarström, H., Forkel, R., Haspelmath, M. y Bank, S. (2021). *Glottolog 4.4*. Max Planck Institute for Evolutionary Anthropology. <https://glottolog.org>.
- Hinton, L. (2014). Orthography wars. En M. Cahill y K. Rice (Eds.), *Developing orthographies for unwritten languages* (pp. 139–168). SIL International.
- Hoeller, A. (1932). *Guarayo-Deutsches Wörterbuch*. Verlag der Missionsprokura der P. P. Franziskaner.
- Kroskrity, P. V. (2004). Language ideologies. En A. Duranti (Ed.), *A companion to linguistic anthropology* (pp. 496–517). Blackwell Publishing.
- Kroskrity, P. V. (2015). Designing a dictionary for an endangered language community: lexicographical deliberations, language ideological clarifications. *Language Documentation & Conservation*, 15, 140–157.
- Leigue Castedo, L. (1957). *El Itenez Salvaje*. Ministerio de Educación.
- List, J. M., Rzymiski, C., Greenhill, S., Schweikhard, N., & Panykh, K., Tjuka, A., Hundt, C. y Forkel, R. (Eds.). (2021). CLLD Concepticon 2.5.0 [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.4911605>
- Lüpke, F. (2009). Data Collection Methods for Field-Based Language Documentation. *Language Documentation and Description*, 6, 53–100.
- Machuqui Dominguez, A. (2021). *Ese Ejjaja Esowijo Ewowi Ja'a Pokiajji: Diccionario Ese Ejja – Español, Español – Ese Ejja*. Foundation for Endangered Languages.

- Majid, A. (2012). A guide to stimulus-based elicitation for semantic categories. En N. Thieberger (Ed.), *The Oxford handbook of linguistic fieldwork* (pp. 54-71). Oxford University Press.
- Marban, P. (1894 [1701]). *Arte de la lengua Moxa con su vocabulario y cathecismo*. B. G. Teubner.
- Payne, T. (1997). *Describing morphosyntax: a guide for field linguists*. Cambridge University Press.
- Peréz Báez, G. y Kaufman, T. (2019). La Ventosa Diidxazá Lexico-Botanical Dictionary. *Dictionaria*, 9, 1-952. <https://dictionaria.clld.org/contributions/diidxaza>.
- Real Academia Española. (2001). *Diccionario de la lengua española*. 22a edición. Editorial Espasa Calpe.
- Reiman, D. W. (2010). Basic oral language documentation. *Language Documentation & Conservation*, 4, 254-268.
- Rice, K. (2018). Reflections on documenting the lexicon. En McDonnell, B., Berez-Kroeker, A. L. y Holton, G. (Eds.), *Reflections on language documentation 20 years after Himmelmann 1998* (pp. 122-131). University of Hawai'i Press.
- van Strekenburg, P. (2003). 'The' dictionary: definition and history. En P. van Strekenburg (Ed.), *A practical guide to lexicography* (pp. 3-17). John Benjamins Publishing Company.