https://doi.org/10.22201/iie.18703062e.2024.124.2861

Diseño de un corpus de la historia del arte y arqueología y generalidades del uso de AntConc para obtener listas de candidatos a términos

Design of an Art History and Archaeology Corpus and an Overview of Using AntConc to Generate Term Candidate Lists

Artículo recibido el 14 de febrero de 2023; devuelto para revisión el 26 de julio de 2023; aceptado el 9 de noviembre de 2023, https://doi.org/10.22201/iie.18703062e.2024.124.2861

Claudio Molina Salinas Universidad Nacional Autónoma de México, Instituto de Investigaciones Estéticas, claudio.molina.salinas@gmail.com, https://orcid.org/0000-0001-5607-9924

Líneas de investigación Lexicografía científica enfocada al patrimonio artístico; documentación del léxico del patrimonio artístico; lenguajes documentales.

Lines of research Scientific lexicography focused on artistic heritage; documentation of lexis of the artistic heritage; (varieties of) documentary language.

Publicación más relevante "Documenting Mexican Folk-Art Linguistic Heritage: The Application of the Sets Theory to Determine Its Common Terminology", *Journal of Ethnic and Cultural Studies* 8, núm. 4 (2021): 238-270, http://www.ejecs.org/index.php/JECS/article/view/811.

Resumen Este artículo se inscribe en el dominio de la lexicografía terminológica (terminografía) y en el de las artes populares. En él se argumenta sobre la necesidad de un corpus de un género textual, muy común en disciplinas como la historia del arte y la arqueología (catálogos de museos), y, naturalmente, se explica su diseño. También se hace una revisión sobre lo que es un gestor de corpus, las herramientas disponibles en línea, en general, y, en particular, se argumenta la razón por la cual se optó por usar AntConc como gestor del corpus, asimismo, se explican las funcionalidades del *software*. Por último, se muestra cómo, al usar técnicas de análisis de corpus en combinación con AntConc, se generan listas de candidatos a términos monoléxicos, se obtienen términos pluriléxicos, y léxico relacionado con los candidatos a términos, con miras a la redacción de definiciones para un diccionario.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Palabras clave Análisis léxico de catálogos de arte; diseño de corpus para las artes; gestores de corpus; análisis textual; extracción semisupervisada de términos.

Abstract The discussion covered by this article falls within the domain of popular arts and crafts, and also that of terminological lexicography (terminography). This work states the need for a corpus pertaining to a textual genre–something that is quite common in areas such as art history and archaeology (museum catalogs)–and explains the design method used. The concept of a corpus management system is also reviewed, and the tools available online are described; in particular, the reason why AntConc was opted for as the corpus management system for this purpose, and the software's characteristics are explained. Lastly, I show how, by using corpus analysis techniques, in combination with AntConc, lists of candidates for monolexical and multilexical terms can be generated, along with vocabulary related to term candidates, for the purpose of writing specialized dictionary definitions.

Keywords Lexical analysis of art catalogs; corpus design for the arts; corpus managers; textual analysis; semi-supervised extraction of terms.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

CLAUDIO MOLINA SALINAS
INSTITUTO DE INVESTIGACIONES ESTÉTICAS, UNAM

Diseño de un corpus de la historia del arte y arqueología

y generalidades del uso de AntConc para obtener listas de candidatos a términos

Introducción

Este artículo es una revisión de nociones provenientes de la lingüística aplicada y ciencias de la información como "corpus", "fuentes de información", "gestor de corpus", "catálogo", entre otras, para que, con base en esta discusión, se ofrezca una visión del diseño conceptual de un corpus de objetos culturales de México que organiza los datos provenientes de catálogos en línea, y de acceso abierto, de museos nacionales y otras instituciones de la memoria y educativas.

La construcción de un corpus como éste representa una iniciativa sin antecedente dentro de la historia del arte y arqueología o lingüística de corpus, que, al integrar información especializada de recursos de uso común en la investigación y difusión del patrimonio cultural (catálogos de bienes artísticos y arqueológicos), pudiera facilitar la recuperación sistematizada de usos contextuales de términos sancionados en catálogos, y, más específicamente, servir como base empírica para la redacción de definiciones de términos.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Corpus lingüísticos y corpus para lexicografía

Un corpus lingüístico (corpus en adelante) es una selección de fragmentos de textos de una lengua o conjunto de textos de materiales escritos y hablados, hoy día casi siempre en formato digital (considerando que muchos textos en la actualidad se producen original o únicamente en este formato) que se ordenan desde criterios lingüísticos explícitos para representar de forma fiel las lenguas naturales y, en consecuencia, realizar ciertos análisis y formalizar generalizaciones sobre la naturaleza de éstas.[1] En la actualidad, los corpus constituyen la base empírica de las investigaciones lingüísticas en general, y, en particular, de la lexicografía.[2]

Tradicionalmente, la representatividad de un corpus, definida en la fase de su diseño, debe considerar las variedades a representar y la proporción en que ocurren éstas en el estado natural de la lengua, nociones conocidas como "variedad" y "equilibrio", respectivamente.[3]

Para el diseño de un corpus también se debe considerar una unidad de muestreo (sean textos completos o fragmentos de éstos), pues ésta determina la utilidad principal del recurso, por ejemplo, un corpus integrado con fragmentos de discurso, párrafos u oraciones aisladas tendría una aplicación concreta para estudios del léxico, ya que, en general, en un párrafo se puede inferir el sentido con el que se usa una palabra; mientras que en un corpus que recoge textos completos se emplea, mayormente, para estudios de un amplio abanico

1. John Sinclair, "Preliminary Recommendations on Text Typology", en *EAGLES (Expert Advisory Group on Language Engineering Standards) EAG-TCWG- CTYP/P* (Pisa: Consiglio Nazionale delle Ricerche/Istituto di Linguistica Computazionale, 1996), 10; John Sinclair, "Corpora for Lexicography", en *A Practical Guide to Lexicography*, ed. Piet van Sterkenburg (Ámsterdam/Filadelfia: John Benjamins Publishing Company, 2003), 167 y Gerardo Eugenio Sierra Martínez, *Introducción a los corpus lingüísticos* (Ciudad de México: Universidad Nacional Autónoma de México-Instituto de Ingeniería, 2017), 4.

2. Marc Kupietz, "Constructing a Corpus", en *The Oxford Handbook of Lexicography*, ed. Philip Durkin (Nueva York: Oxford University Press, 2017), 62.

3. Para ampliar sobre esto véase: Douglas Biber, Susan Conrad y Randi Reppen, *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge: Cambridge University Press, 1998), 243-250; Gerardo Eugenio Sierra Martínez, "Diseño de *corpus* textuales para fines lingüísticos", en *Memorias del IX Encuentro de Lingüística en el Noroeste*, t. II, eds. Zarina Estrada Fernández, Ana Lidia Munguía Duarte y Rosa María Ortiz Ciscomani (Hermosillo: Universidad de Sonora, 2008), 450-454 y Kupietz, "Constructing a Corpus", 63-68.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

de aspectos lingüísticos.[4] En resumen, se asume que la representatividad de un corpus se define a partir de la variedad y el equilibrio, y cómo ambos reflejan la población estadística de la lengua o variantes representadas en él; mientras que, a partir de la unidad de muestreo, se define la aplicación del corpus a cierto fenómeno lingüístico.

En general, se asume que la representatividad de un corpus se modela estadísticamente dependiendo de la realidad que se quiere describir. Para ello se consideran aspectos como el origen de los datos, los informantes, las temáticas, los tipos textuales, la temporalidad de los datos, entre otros.[5] Sin embargo, existe un enfoque más o menos novedoso en el que se favorece la inclusión del mayor número de materiales escritos y hablados posibles, de todas las variedades disponibles, pues se acepta que la representatividad de un corpus varía en función del dominio a investigar, y del propósito o pregunta de investigación del analista que lo utiliza. Podría decirse que, desde esta óptica, se privilegia la amplitud del corpus en detrimento de la representatividad (explicada antes), pero no es así del todo. Lo que se asume es que un analista, al usar el corpus en una investigación específica, recuperará submuestras de éste (corpus virtuales) y modelará la representatividad mediante búsquedas específicas (en el caso de corpus digitales);[6] esto hace a los corpus diseñados con este enfoque proyectos económicos, reusables y polivalentes.[7]

Aparte de la representatividad de un corpus, hay que considerar también que éstos se pueden clasificar en diferentes tipos; sin embargo, considerando el alcance de este documento, sólo ahondaré en la caracterización de una clase muy específica: los corpus para lexicografía.[8]

4. Douglas Biber, "Representativeness in Corpus Design", *Literary and Linguistic Computing* 8, núm. 4 (octubre de 1993): 243-244; Joan Torruella y Joaquim Llisterri, "Diseño de *corpus* textuales y orales", en *Filología e informática. Nuevas tecnologías en los estudios lingüísticos*, ed. José Manuel Blecua (Barcelona: Universidad Autónoma de Barcelona /Editorial Milenio, 1999), 60-62.

5. Sinclair, "Preliminary Recommendations on Text Typology", 11-5 y Sierra Martínez, *Introducción a los corpus lingüísticos*, 49-57.

6. Marc Kupietz, Cyril Belica, Holger Keibel y Andreas Witt, "The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research", en *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*, ed. Nicola Calzolari (Malta: European Language Resources Association [ELRA], 2010), 1849 y Kupietz, "Constructing a Corpus", 66-67.

7. Kupietz, "Constructing a Corpus", 66.

8. Para conocer más sobre las tipologías de corpus existentes, se recomienda revisar a John

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Un corpus para lexicografía es un recurso pensado y diseñado *ex professo* para ser el centro de un proyecto de diccionario.[9] En general, el desafío de quien diseña un corpus para lexicografía radica en comprender la riqueza de los materiales que se incluirán en él, pues, si no se hiciera, podría haber sesgos en cuanto al condicionamiento de aparición de ciertas unidades léxicas, relacionado con algunas temáticas de los textos; la sobrerrepresentación del léxico usado por un autor; o la falta de documentación de algunas unidades léxicas de uso común. Comprender esta riqueza de los materiales y considerarla en el diseño del corpus garantiza documentar la mayor cantidad de léxico de una lengua, asumiendo que éste será el punto de referencia para documentar los significados globales en uso del léxico y que, por su diseño, representaría el número más elevado de vocablos posibles.[10] Con esto en mente, en los apartados siguientes se caracterizará el "Corpus de Catálogos de Bienes Culturales de México" y más adelante se ofrecerán algunas aplicaciones de éste a la investigación de las terminologías del dominio y la definición de sus términos.

Visión general del corpus

Antes de comenzar un proyecto de creación de un corpus conviene asegurarse de que no existen iniciativas semejantes, pues algunas veces podría ser más razonable y mucho más valioso reutilizar o mejorar un corpus existente, que construir uno nuevo. Así, para evitar la creación de trabajos redundantes, se recomienda corroborar la existencia o no de proyectos semejantes, consultando índices como The Catalog, el LRE map, el catálogo del Linguistic Data Consortium y el Virtual Language Observatory,[11] y hacer búsquedas en catálogos de bibliotecas de universidades y recursos de centros de investigación locales. En este caso, como se ha comprobado que no existe un corpus lexicográfico de catálogos del dominio de la historia del arte y arqueología en México, se presenta la visión general de un corpus como éste y algunas de sus características

Sinclair, "Preliminary Recommendations on Text Typology"; Torruella y Llisterri, "Diseño de corpus textuales y orales" y Sierra Martínez, *Introducción a los* corpus *lingüísticos*, entre otros.

9. Sinclair, "Corpora for Lexicography", 167.

10. Luis Fernando Lara Ramos y Roberto Ham Chande, "Base estadística del Diccionario del Español de México", en *Investigaciones lingüísticas en lexicografía*, ed. Luis Fernando Lara Ramos (Ciudad de México: El Colegio de México, 1979), 13-14, 17.

11. Kupietz, "Constructing a Corpus", 63.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

destacables, poniendo particular atención al diseño lingüístico del corpus (objetivo, selección de la documentación y procesamiento de ésta).

Caracterización del corpus (objetivos generales de éste)

El objetivo general de esta propuesta es crear un corpus para lexicografía, sincrónico y monolingüe, basado en catálogos de bienes culturales de México disponibles en línea; que, desde un punto de vista lexicográfico, facilite la investigación respecto a las terminologías del dominio; y que aporte pistas sobre el significado de los términos y las tradiciones de uso de éstos para describir bienes culturales mexicanos.

Fuentes de información, selección de la documentación y representatividad del corpus

El término *fuente de información* denomina a todo material o producto, original o reelaborado, que pueda aportar información o ser usado como un testimonio para construir conocimiento.[12] Desde el punto de vista de las ciencias documentales, las fuentes de información se organizan en tres niveles: primarias, secundarias y terciarias. Las fuentes primarias constituyen información original, resultado de una investigación o de una actividad innovadora y creativa, y su función principal es comunicar los resultados del conocimiento; se estructuran como discursos textuales coherentes, consecutivos y dependientes en su significado. Algunos ejemplos son libros, tesis, artículos, correspondencia, entre otros.[13]

Por otra parte, las fuentes secundarias "contienen información primaria reelaborada, sintetizada y reorganizada, o remiten a ella. Son especialmente diseñadas para facilitar y maximizar el acceso a las fuentes primarias o a sus contenidos. Se estructuran en discursos textuales o icónicos fragmentados, coherentes e independientes en su significado, y siguen la lógica y la estructura de

12. Isabel de Torres Ramírez, "Las fuentes de información. Metodología del repertorio bibliográfico", en *Manual de Ciencias de la Documentación*, ed. José López Yepes (Madrid: Pirámide, 2002), 317.

13. Susana Romanos de Tiratel, *Guía de fuentes de información especializadas. Humanidades y Ciencias Sociales* (Buenos Aires: GREBYG/Centro de Estudios y Desarrollo Profesional en Bibliotecología y Documentación, 2000), 18 y 20.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

las bases de datos".[14] Los catálogos, diccionarios, enciclopedias, bibliografías, índices y bases de datos bibliográficos son ejemplos de fuentes secundarias.[15]

Por último, las fuentes terciarias contienen información sobre las secundarias y remiten a ellas, por ejemplo: las guías de obras de referencia o las bibliografías de bibliografías.[16]

Hay que considerar que, en un sentido amplio, también son fuentes de información cualquier otro elemento disponible que contenga símbolos con la capacidad de significar, independientemente del soporte en el que se encuentren.[17] Esto implica que toda huella, vestigio, testimonio, resto biológico, monumento, obra de arte o, incluso, los restos encontrados en un yacimiento arqueológico son, también, fuentes de información.[18] Ya que en áreas disciplinares como la historia del arte y la arqueología se recurre tanto a fuentes textuales (libros, artículos, reportes, etcétera), como a la consulta directa de fuentes de información como fotografías, pinturas, esculturas, grabados, monolitos, cerámicas, restos óseos, entre otros, esta última definición de "fuente de información", en sentido amplio, resulta pertinente.

Para disciplinas como la historia del arte y la arqueología, la relación entre las fuentes primarias, que funcionan como documentos[19] "no textuales" y que no aportan información de forma escrita, pero sí de forma simbólica, y los catálogos[20] que los describen es de gran relevancia. Pensemos que en esta relación biunívoca el "catálogo es, en cierta medida, un sustituto de la cosa"[21] que, para un observador o investigador lego, son la mejor o la única alternativa para aproximarse a un primer análisis.

14. Romanos de Tiratel, *Guía de fuentes de información especializadas*, 19.

15. Romanos de Tiratel, *Guía de fuentes de información especializadas*, 21.

16. Romanos de Tiratel, *Guía de fuentes de información especializadas*, 19 y 22.

17. Romanos de Tiratel, *Guía de fuentes de información especializadas*, 5.

18. Torres Ramírez, "Las fuentes de información", 317.

19. Aquí tomo el sentido amplio del término *documento*, propuesto por Suzanne Briet, *¿Qué es la documentación?* (Santa Fe: Universidad Nacional del Litoral, 1960), 11, y asumo que un documento es todo signo concreto o simbólico, conservado o registrado, destinado a representar, reconstruir o demostrar un fenómeno físico o conceptual. Por tanto, considero fotografías, pinturas, esculturas, así como libros, tesis o artículos, documentos.

20. Los catálogos son obras razonadas, muy bien caracterizadas, siempre de carácter textual, que describen las particularidades de las cosas y las clasifican bajo ciertos principios de orden. Véase Paul Otlet, *El tratado de documentación. El libro sobre el libro. Teoría y práctica* (Murcia: Editum, 2007), 170.

21. Otlet, *El tratado de documentación,* 287.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Hasta ahora he argumentado sobre la relevancia que tienen los catálogos para disciplinas como la historia del arte y la arqueología –esto a la luz de su objeto de estudio y su tradición– y cómo éstos constituyen parte fundamental de la documentación del dominio. Además, ya se ha señalado que no existe un corpus lexicográfico formado a partir de este tipo de documentación; sin embargo, aún faltaría argumentar sobre cuáles pudieran ser los catálogos que se integrarían en el corpus y las razones de esta decisión, así como ahondar en la caracterización de la representatividad de esta selección.

Un posible punto de partida para integrar este corpus son 34 catálogos que recogen, en total, más de 80 000 bienes de interés cultural de México, descritos de manera minuciosa; que están en mi poder y puedo usar con libertad para esta propuesta: tres de ellos refieren al patrimonio de la Universidad Nacional Autónoma de México (UNAM), y son propiedad del Instituto de Investigaciones Estéticas (IIE); otros 30 son documentos que he solicitado formalmente a la Secretaría de Cultura de México[22] (SC); y uno más de la UNAM,[23] la mayoría de estos catálogos están en línea, excepto los que son propiedad del IIE.

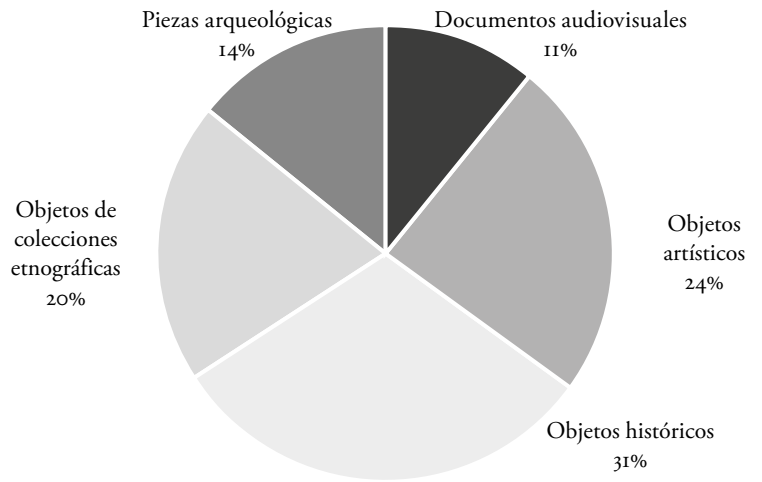
Estos 34 catálogos describen documentos audiovisuales (grabaciones de audio y video) (11%), objetos artísticos (pinturas, esculturas, grabados, entre otros) (24%), objetos históricos (31%), piezas arqueológicas (14%) y objetos de colecciones etnográficas (20%), todos ellos provenientes de museos o instituciones como el Museo Nacional de Culturas Populares, la Fototeca Nacional, el Museo Nacional de San Carlos, la UNAM, entre otros. En la figura 1 se puede ver representada esta proporción.

Hacer una preselección de los catálogos, en su totalidad, o de los objetos que representan, en lo particular, y modelar la representatividad del corpus es la alternativa más económica en cuanto a la optimización del trabajo de selección y procesamiento documental, pues una muestra heterogénea y acotada que recoja las variedades de objetos en la proporción estadística que ocurren para el contexto nacional mexicano sería suficiente para garantizar la representatividad del recurso. Lamentablemente no existe un trabajo cuantitativo conocido sobre el total de la población de bienes culturales de México, por tanto, definir la proporción y variedad de objetos descritos que deberían incluirse en el corpus resulta, en este momento, imposible. Considerando esto,

22. Información que me fue entregada mediante una petición formal respondida con el oficio DGTIC/0013/2020, del día 21 de enero de 2020.

23. Disponible en http://www.patrimonio.unam.mx/BAC/ (consultado por última vez el 11 de marzo de 2024).

https://doi.org/10.22201/iie.18703062e.2024.124.2861



1. Porcentaje de los tipos de objetos en los 34 catálogos.

la alternativa plausible para la conformación del corpus es tomar todos los registros disponibles e integrarlos y, a la vez, definir condiciones para que el investigador-analista seleccione las variedades de objetos que le interesa estudiar (en el apartado "Aplicación de cinco funcionalidades de AntConc a la investigación terminográfica" se explicará cómo se prevé que se lleve a cabo este procedimiento). En este caso, el corpus podría considerarse una muestra primordial, tal y como se ha explicado antes.

Por otra parte, es indiscutible que utilizar un tipo textual único implica un sesgo; sin embargo, por el enfoque lexicográfico del corpus y la intención de documentar la mayor cantidad de términos del dominio, éste resulta conveniente en una primera etapa de desarrollo del corpus, pues la mayor parte de la información recogida en los catálogos denomina con términos a tipos de objetos, su materialidad, las técnicas empleadas en su manufactura, su estilo, los posibles deterioros asociados a éstos, su lugar de origen, entre otros.

En suma, la idea de integrar un corpus de catálogos se justifica porque, para la historia del arte y la arqueología, el género textual describe minuciosamente fuentes o documentos "no textuales" con un tipo de fuente que verbaliza, mediante el uso de términos descriptivos, las características formales, funciones, contextos de origen y otros datos relacionados con los objetos. Esta característica particular de los catálogos del dominio se alinea en forma conveniente con los objetivos de un corpus lexicográfico como el propuesto.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Procesamiento de la documentación

En el contexto museal mexicano han existido múltiples enfoques respecto a cómo estructurar las fichas de objetos patrimoniales y sobre cómo almacenar sus catálogos, esto último de interés para esta discusión. Por ejemplo, podemos encontrar desde catálogos en papel, catálogos almacenados en tarjetas perforadas, que, me permito aclarar, funcionan más como una pieza de historia o de museo, y no son funcionales actualmente; hasta catálogos que se encuentran en un formato de bases de datos de Microsoft Access o, incluso, en XML, pasando por archivos de Word, PDF y los muy comunes catálogos en Excel.

Estos catálogos, así como la mayoría de las fuentes de información secundaria, sin importar el formato físico o digital en el que se encuentren, se estructuran como discursos fragmentados, y suelen adquirir la forma de bases de datos, apegados a un modelo específico. Para Van Hooland y Verborgh hay cuatro modelos de datos utilizados principalmente: uno tabular, uno relacional, uno *meta-markup language* (quizá pueda denominarse también "modelo jerárquico") y uno RDF (del inglés Resource Description Framework), independientemente del formato en el que se almacenen. Los autores explican que, en un modelo tabular, cada ítem se describe con una línea de valores organizados en campos (o columnas), tal como se ve en una hoja de cálculo, incluso, la primera línea puede usarse para definir el nombre de cada campo; para un modelo relacional, los datos también se estructuran en tablas, pero éstas tienen el potencial de relacionarse o conectarse, a su vez, con otras tablas; en el caso de un modelo *meta-markup language*, los datos se organizan de forma jerárquica, desde un elemento raíz o base, lo que les da una "aparente" figura de árbol; mientras que, en el RDF, los datos de cada objeto o ítem se expresan como una tripleta, que conecta a un sujeto (dato) con un objeto (otro dato) mediante una relación que emula oraciones simples de cualquier lengua natural.[24]

Los 34 catálogos antes mencionados se encuentran en formato de Office Excel, lo que implica que sus datos se organizan bajo un modelo tabular. Desde la perspectiva de esta propuesta, esto es conveniente, ya que el manejo de los documentos resulta relativamente sencillo para perfiles con poco entrenamiento informático. Además, es natural que el corpus se mantenga en

24. Seth van Hooland y Ruben Verborgh, *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata* (Londres: Facet Publishing, 2014), 16.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

formato electrónico, sin embargo, hay que pensar que los datos de cada uno de los catálogos tienen que pasar por un proceso de adecuación para poder integrarlos y utilizarlos; esto es necesario para que el *software* de gestión de corpus pueda manejarlos.

El procesamiento de la documentación y la preparación de los catálogos para que funcionen con el gestor de corpus debe considerar, primero, la selección de la información de los datos lingüísticamente relevantes; segundo, adoptar un procedimiento de preservación de la información de los catálogos; y tercero, la transformación de la forma en la que se presenta cada ítem para que sea más legible en un gestor de corpus.

Sobre la selección de la información de los datos lingüísticamente relevantes hay que considerar el principio de no intervenir los datos originales de los catálogos, o modificarlos lo menos posible. Sin embargo, también hay que asumir que, para hacer más efectiva la búsqueda de información en un corpus, conviene cuidar que los textos se ajusten a las convenciones ortográficas de la comunidad que va a usarlo; así, cualquier analista podría consultarlo y leer fragmentos de éste sin correr el riesgo de hacer interpretaciones equivocadas, asociadas a diferencias en la transliteración del corpus,[25] lo que implica una inevitable manipulación (ligera) de las fuentes.

Al considerar que los 34 catálogos que integrarán el corpus se produjeron originalmente en español contemporáneo y por hablantes de esta lengua, se puede asumir que estas fuentes de información se ajustan a las convenciones ortográficas del español y, por ende, no es necesario hacer modificaciones en cuanto a la ortografía de los datos de las fuentes o, incluso, optar por la lematización de sus términos.

Si es cierto que el enfoque para esta propuesta implica intervenir lo menos posible los datos de los catálogos, sí considero necesario seleccionar los elementos que proporcionan información lexicográficamente relevante y descartar elementos de información administrativa cuya difusión sea delicada. Por una parte, el criterio seguido para elegir los elementos de información que aportan datos de relevancia lexicográfica es revisar cuáles son los elementos que describen a los ítems con un término, o con una frase o párrafos susceptibles de incluirlos –por ejemplo, los campos: "Tipo de objeto", "Materia prima", "Técnica de manufactura", "Estado de conservación", "Observaciones", entre otros– y organizarlos desde una lógica de prelación.

25. Sinclair, "Corpus Processing", 183.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

En cambio, las informaciones administrativas y de difusión cerrada al público deben eliminarse del corpus: por un lado, hay datos administrativos que no aportan información lexicográficamente relevante, pero sí generan ruido en el análisis del corpus (por ejemplo, los diferentes números de registro que han tenido los objetos, el registro de movimientos que ha tenido la obra o el número de imágenes disponibles del objeto); o, por otra parte, información delicada que no debería publicarse o que no conviene que sea del dominio público (como la ubicación del objeto dentro de las bóvedas de un museo o su valuación).

Sobre la preservación de los archivos de un corpus, Sinclair[26] recomienda almacenarlos en archivos de texto plano (.txt), pues es un formato genérico en el que no suelen corromperse los contenidos, en oposición a otros tipos de archivos en los que las actualizaciones pueden comprometer la integridad de los datos; por ello debería adoptarse esta recomendación técnica.[27] En cuanto al cambio de formato y la forma en que se presenta cada ítem en el catálogo, hay que considerar, antes de convertir archivos de Office Excel a texto plano, que el manejo de los catálogos en este formato podría dificultar la lectura de contenidos de datos originalmente organizados en un modelo tabular, pues esta conversión descontextualiza la información.

Al revisar de manera cuidadosa la variedad de datos mostrados en la figura 2, elaborada a partir del primer registro de la base de datos del Museo Nacional de Culturas Populares (MNCP), está claro que hay algunos casos de los que se puede inferir el campo de información que se registra, por ejemplo: "Máscara" corresponde al descriptor "tipo de objeto", o "Municipio de Mitontic, Chiapas" a la "Procedencia del objeto"; sin embargo, hay casos en los que no son tan evidentes estas relaciones, "Máscaras" o "Personaje de carnaval", que son parte de los elementos de información relacionados a "Colección" y "Observación", respectivamente.

Por ello se plantea la transformación de los catálogos, no sólo en el plano del tipo de archivo, sino también en la forma en que se presentan los datos de

26. Sinclair, "Corpus Processing", 183.

27. Aceptar sin mayor discusión que el formato adecuado para almacenar y preservar un corpus es el texto plano (0.txt) podría extrañar al que leyere este artículo, ya que no se han discutido otras alternativas como el formato CSV (Comma Sparated Values) que sirve muy bien para vincular información de Exceles, pero, como se verá más adelante, la herramienta de gestión de corpus utilizada en el proyecto sólo puede leer archivos de texto plano. Ésa es la razón por la que se concluye que este formato es el más adecuado para el proyecto, sin mayor discusión al respecto.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Frecuencia absoluta	Valor MI	Término
1861	3.77495	madera
603	3.94422	pintura
60	4.37354	licra
54	3.15115	piel
50	2.47824	crin
48	2.54911	pelo
47	2.99739	cuero
44	2.92608	papel
40	2.42868	cuerno
34	3.41661	barniz
32	3.61866	cedro
32	3.56619	barro
26	2,82606	tela
24	2,82922	cartón
22	2.54747	ixtle

2. Materiales relacionados a "máscara".

cada ítem; para esto se propone arreglar los datos como párrafos (delimitados entre "puntos y aparte") y especificar, para cada uno de los registros, el campo de información al que está relacionado el término o términos descriptivos. Enseguida se presenta la forma en la que se reelaboraría este primer registro conservando los datos lingüísticamente relevantes, ordenándolos de forma alfabética, cambiando a formato de párrafo la información y agregando los descriptores de cada campo.

Este procedimiento mantiene el contexto original del catálogo y le brinda un "sentido completo" a cada término usado en él, pero en un nuevo formato, al tiempo que prepara los textos para su preservación y explotación con un gestor de corpus. Además, hay que aclarar que se han considerado dos elementos de información que podrían funcionar para el trabajo con el corpus, pese a no aportar información lexicográficamente relevante, pero sí de valor documental: el elemento de información respecto a "quién elaboró el registro" y el "número de registro" del objeto. El elemento de información: "quién elaboró el registro" servirá para controlar la variabilidad terminológi-

https://doi.org/10.22201/iie.18703062e.2024.124.2861

ca; y el "número de registro", para referir al objeto, en la fuente original en la que se documenta. A continuación se ahonda de qué manera se gestionarán estos archivos y qué funciones de análisis textual se pueden aplicar al corpus en este formato.

Sistemas de gestión de corpus, aplicaciones de análisis textual y AntConc

Los sistemas de gestión de corpus (gestores de corpus, en adelante) son aplicaciones muy especializadas que permiten trabajar con grandes volúmenes de información en formato de texto, ejecutar consultas relacionadas con éstos y convertir colecciones textuales en corpus digitales.[28] Estas herramientas permiten hacer análisis cuantitativos, cualitativos, y administrar mejor la documentación de los corpus. En general, los análisis cuantitativos derivados del uso de estos gestores se relacionan con el cálculo de datos estadísticos de un corpus, los análisis cualitativos se ciñen a la creación de listas de léxico, la obtención de colocaciones o recuperación de N-gramas, la recuperación de contextos de uso (búsquedas de concordancias), entre otros; y, sobre la administración de la información, estos gestores permiten el manejo de los archivos de un corpus, el procesamiento de los textos y administración de metadatos relacionados con ellos (autor, género, tema…).[29]

En internet hay, al menos, diez gestores de corpus disponibles: AntConc, Sketch Engine, WordSmith Tools, The IMS Open Corpus Workbench, GECO 3, Corpógrafo V.5, Linguakit, Concordance, MonoConc y TextStat; sin embargo, no todos los gestores de corpus cuentan con características o funcionalidades útiles para el trabajo lexicográfico que se plantea; por ejemplo, de estas diez herramientas, sólo AntConc, Sketch Engine y WordSmith Tools permiten hacer listas de unidades léxicas, recuperar contextos de uso (hacer búsquedas de concordancias), obtener colocaciones, enlistar N-gramas o crear listados de palabras clave, razón por la que se descarta usar los otros siete ges-

28. Gerardo Eugenio Sierra Martínez, Julián Solórzano Soto y Arturo Curiel Díaz, "GECO, un Gestor de Corpus colaborativo basado en web", *Linguamática* 9, núm. 2 (2018), 57.

29. Sierra Martínez, Solórzano Soto y Curiel Díaz, "GECO, un Gestor de Corpus colaborativo basado en web", 57 y George Kouklakis, George K. Mikros, George Markopoulos e Ilias Koutsis, "Corpus Manager: A Tool for Multilingual Corpus Analysis", en Corpus *Linguistics Conference 2007* (Birmingham: University of Birmingham, 2007), 3-4.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

tores. Adicionalmente, AntConc, Sketchengine y WordSmith Tools permiten la anotación gramatical del corpus y otras funcionalidades avanzadas.[30] Estas condiciones hacen a estas tres herramientas candidatas a ser usadas en la gestión del corpus propuesto.

AntConc es un *software* amigable para el usuario, multiplataforma, con una interfaz fácil de usar, en constante desarrollo y actualización, ideal para corpus pequeños o medianos[31] y, con base en mi experiencia utilizando las tres herramientas, me permito afirmar que es tan efectivo como los otros dos gestores; sin embargo, AntConc destaca porque es *freeware*, o de licencia libre, y no necesita instalarse. Esto lo convierte en la mejor alternativa para compartir, manipular o transportar un corpus como el que se plantea. Por ello se propone usar esta herramienta para administrar y explotar el recurso, cuando menos en una primera etapa de trabajo.

Quizás un inconveniente de AntConc, respecto a los otros dos gestores, es que no está pensado para la publicación de corpus en línea, sin embargo, esto no representa un problema ahora, ya que legalmente puedo usar los catálogos que conformarían el corpus, aunque aún no cuente con los permisos respectivos para publicarlo en internet, por ejemplo.

Las funcionalidades con las que cuenta AntConc, en su versión 3.4.4 y subsecuentes, son siete: *Concordance Tool* o "Función de concordancias", *Concordance Plot Tool* o "Diagrama de concordancias", *File View Tool* o "Función de vista en el archivo", *Clusters/N-grams*, *Collocate* o "Colocaciones", *Word List* o "Lista de palabras" y *Keyword List* o "Lista de palabras clave".[32] No todas estas funcionalidades resultan relevantes para el enfoque del corpus lexicográfico que se ha planteado, por tanto, en el siguiente apartado se ahondará en la definición de las que sí lo son para esta propuesta, es decir: *Word List*, *Clusters/N-grams*, *Collocate*, *Concordance Tool* y *File View Tool*, y se esbozarán algunas aplicaciones de cada una de ellas.

30. Sierra Martínez, Solórzano Soto y Curiel Díaz, "GECO, un Gestor de Corpus colaborativo basado en web" y Kouklakis, Mikros, Markopoulos y Koutsis, "Corpus Manager: A Tool for Multilingual Corpus Analysis", 57-8.

31. Laurence Anthony, *Laurence Anthony's AntConc*, https://www.laurenceanthony.net/software/antconc/ (consultado por última vez el 11 de marzo de 2024).

32. Jesús Aparicio Boussif, *AntConc (Windows, Macintosh OS X y Linux) versión 3.4.4 (in Spanish)* (Córdoba: Universidad de Córdoba, 2014), 2-3.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Aplicación de cinco funcionalidades de AntConc a la investigación terminográfica

AntConc permite aproximarnos al análisis de un corpus desde el punto de vista de la recurrencia de las formas (creando listas de unidades léxicas de alta repetición, descubriendo unidades pluriléxicas recurrentes en un texto e identificando unidades léxicas que coocurren con términos ya identificados) y desde el punto de vista de la documentación del uso del léxico en contexto. En seguida veremos cuáles son estas funcionalidades.

El primer paso antes de comenzar el análisis textual con el gestor de corpus es seleccionar y cargar en la herramienta uno o más archivos con los que se quiera trabajar (catálogos del corpus). En este momento, el investigador-analista seleccionará las variedades de objetos de su interés, escogiendo los catálogos que más le convengan. Tras esta operación, se puede comenzar con el análisis.

Probablemente, la funcionalidad más importante y, quizás, el primer paso de todo análisis de corpus es, desde una perspectiva lexicográfica, crear listados de unidades léxicas. Para ello, AntConc tiene la funcionalidad *Word List*, que permite presentar todas las unidades monoléxicas de un corpus en forma de lista ordenada por frecuencia (ascendente o descendente) o por orden alfabético (canónico o inverso); esta funcionalidad se puede configurar para considerar o ignorar las mayúsculas y minúsculas en su análisis.[33] Además, las versiones más recientes de AntConc (3.5.9 y siguientes) pueden hacer un filtrado de los resultados a partir de listas de exclusión de unidades léxicas que el analista quiera que no se consideren en los resultados.[34]

Trabajar con la funcionalidad "lista de palabras" de AntConc es la primera aproximación al análisis propuesto, ya que, primero, desde el punto de vista lingüístico, es fundamental conocer todas las unidades monoléxicas que hay en el corpus y filtrar las que, de antemano, se sabe que no interesarán para el trabajo terminográfico (por ejemplo, si se quisieran excluir de los resultados los artículos, preposiciones, conjunciones, interjecciones, entre otras); y segundo, desde el punto de vista de la herramienta, las "listas de palabras" indizan el léxico dentro del corpus y, por ende, ninguna otra funcionalidad se puede usar si no se ha hecho dicha lista.

33. Boussif, *AntConc*, 7.
34. Anthony, *Laurence Anthony's AntConc*.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Hay que considerar que la lista generada por *Word List* sólo incluye unidades monoléxicas, como "mármol", "talla" o "fotografía", pero es muy común que en el dominio haya términos pluriléxicos como "piedra caliza", "óleo sobre tela" y "fotografía blanco y negro", por lo que hay dos funcionalidades de AntConc que se pueden usar para identificar términos pluriléxicos: *Clusters/N-grams* y *Collocate*.

La función *Clusters/N-grams* de AntConc sirve para hacer listados de términos pluriléxicos, mediante el análisis del corpus en su totalidad; y busca grupos de términos de longitud 'N', a partir de expresiones comunes.[35] Esta funcionalidad crea listados de candidatos a términos pluriléxicos cuya longitud se intuye, pero cuyos formantes se desconocen, basándose en un método de descubrimiento que arroja listados de candidatos a términos que deben cotejarse contra algún control terminológico o de autoridad.

Por otra parte, la función *Collocate* se "utiliza para generar una lista ordenada de las colocaciones que aparecen junto al término buscado, permitiendo encontrar patrones".[36] Con esta funcionalidad de AntConc se pueden identificar posibles complementos asociados a un término conocido; es decir, el analista puede definir un término base de búsqueda: "óleo", y la herramienta le propondría como término: "óleo sobre tela", entre otros resultados.[37] La combinación del uso de las tres herramientas generaría tres listados de candidatos a términos: el primero, de unidades monoléxicas, y el segundo y tercero, de unidades pluriléxicas.

En conjunto, las tres funcionalidades permitirían aproximarse a la identificación de los términos usados en los catálogos; para ello se necesita un proceso de cotejo contra otras fuentes de información, y corroborar que son términos del dominio. Inevitablemente, el analista debe hacer este trabajo de revisión del léxico en índices, tesauros u otros diccionarios (controles terminológicos y de autoridad), o con especialistas del área.

35. Boussif, *AntConc*, 5-6.

36. Boussif, *AntConc*, 2, 6.

37. Una de las medidas que utiliza la herramienta para estimar estas colocaciones, incluso es la que está seleccionada por *default*, es la información mutua (MI, por sus siglas en inglés), que esencialmente mide la dependencia estadística de aparición de dos ítems léxicos en el corpus analizado. Otras medidas que se pueden usar son *Log-likelihood*, *MI* + *Log-likelihood*, y *T-Score*, naturalmente, esto dependerá de las necesidades del analista y del entendimiento de estas medidas por su parte. En este caso siempre se usó *MI*.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Las funcionalidades de AntConc que permiten documentar los términos en su contexto de uso son dos: *Concordance Tool* y *File View Tool*, en combinación. La funcionalidad *Concordance Tool* "muestra los resultados de búsqueda en formato KWIC (palabras clave en contexto), permitiendo observar cómo se usan […] en un corpus de textos".[38] Los resultados de las búsquedas se presentan en forma de listado, y se acompañan de un número de contexto (asignado automáticamente) y la referencia a la fuente de la que se extrajo. Hay que aclarar que estas búsquedas también se pueden construir usando expresiones regulares como: [*] usada para recuperar cero o más caracteres; [+], para cero o sólo un carácter; [¿] para cualquier carácter; [@] para cero o cualquier palabra; [#] para cualquier palabra; [|] para buscar concordancias de un término A u otro término B, que podrían no coocurrir; y, por último, [&] que no haya más palabras.

Una aplicación terminográfica de esta herramienta es construir un fichero terminológico, es decir, guardar, en archivos independientes, los resultados de las búsquedas de los contextos de uso de los términos que se consulten. A partir de esta documentación, un analista o equipo de trabajo podrían intentar inferir algunas condiciones que caracterizan a sus términos y, con esto, ir creando definiciones de ellos.

Al final, si el contexto de uso fuera poco claro, y el analista quisiera ahondar mucho más ampliamente en el texto, la funcionalidad *File View Tool* le permite examinar con más detalle los resultados generados en *Concordance Tool*,[39] llevando al analista al documento completo, a la localización exacta del contexto extraído, para que pueda revisar y aclarar las dudas que considere, dando un clic sobre la concordancia.

Identificación de candidatos a términos e inferencia del léxico que podría servir para definir la materialidad o manufactura de los tipos de objetos

En esta sección se explica cómo seleccionar un subcorpus, y se ofrecen los resultados de la aplicación de las funcionalidades descritas en la sección inmediata anterior, es decir, resultados de la creación de una lista de candidatos a términos monoléxicos, la búsqueda de términos pluriléxicos que incluyen una

38. Boussif, *AntConc*, 2 y 3.
39. Boussif, *AntConc*, 3-4.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

base léxica específica y de los que sólo se conoce su número de formantes; y, al final, se ejemplifica la recuperación de términos relativos a la materialidad y técnicas de manufactura que funcionarían como elementos definitorios a considerar para el término "máscara".

Integración del subcorpus para su estudio

Ya que se tiene un interés relacionado con el estudio de la terminología del arte popular en México y la creación de un glosario de términos de este dominio, de los 34 catálogos en mi poder, se integraron en un subcorpus: el catálogo del Museo Nacional de Culturas Populares y el del Museo Nacional de Antropología (sección de etnografía) (MNA), ya que en otros catálogos no es común que se registren objetos relacionados con el arte popular en México.

Este subcorpus contiene 1 739 870 ocurrencias (*word tokens*) y 21 502 tipos (*word types*),[40] en total. En términos generales, este subcorpus se conforma a partir de 61 448 registros de bienes de interés cultural relacionados al arte popular en México, de los cuales, 46 775 son registros del MNA (76.1% del total) y 14 673 (23.9% del total) provienen del MNCP.

Obtención de candidatos a términos con la funcionalidad Word List

Para obtener una lista de candidatos a términos se usó la herramienta *Word List* o "Lista de palabras". El resultado de la aplicación de esta funcionalidad al corpus generó una primera lista, de la que no todos los elementos se pueden considerar términos, ya que *Word List* cuenta todas las formas ortográficas que se encuentran entre espacios, signos de puntuación o cualesquiera combinaciones posibles, incluyendo en los resultados palabras funcionales, símbolos, números, letras aisladas, entre otras.

Al considerar estos resultados, esta primera lista requirió dos pasos más para mejorar su precisión. Estos procedimientos son: utilizar una lista de fil-

40. Aquí se debe entender que una "ocurrencia" es cada una de las apariciones de una palabra en un texto; mientras que un "tipo" es cada una de las palabras encontradas, eliminando de esta cuantificación sus repeticiones, véase Luis Fernando Lara Ramos, *Curso de Lexicología* (Ciudad de México: El Colegio de México, 2006), 155-156, es decir, si en un catálogo hay 100 ocurrencias del término *máscara*, en este ejemplo, sólo se cuantificaría un tipo.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

trado (*stoplist*) y, después, aplicar a los resultados reglas lingüísticas de lematización. En principio, se hizo una nueva búsqueda utilizando una lista de filtrado que excluye 3 952 tipos que nunca serían términos del dominio, la cual he ido integrando a lo largo de los últimos ocho años de trabajo;[41] además, en esta lista de filtrado incluí los descriptores que se agregaron a los catálogos, a saber: "Número de registro", "Tipo de objeto", "Materia prima", "Técnica de manufactura", etcétera. En segunda instancia, se aplicaron a los resultados reglas lingüísticas de lematización.[42] Los primeros diez resultados de este procedimiento se pueden ver en la figura 3.

Ranking	Frecuencia	Candidato a término	Formas lematizadas
1	14984	Estado	
2	11732	madera	madera 11724 maderas 8
3	11050	algodón	
4	10961	México	
5	10738	doméstico	domésticos 10738
6	10737	enser	enseres 10737
7	10209	indumentaria	
8	9316	Oaxaca	
9	7769	barro	
10	7744	pintura	pintura 6317 pinturas 1427

3. Primeros 10 resultados después del filtrado y lematización en la funcionalidad *Word List*.

41. En esta lista de filtrado se incluyen pronombres, números, preposiciones, artículos, interjecciones, adverbios y conjunciones, verbos de semántica ligera (como *hacer*) que, por su naturaleza gramatical y semántica, no serían términos de este dominio u otras áreas del saber humano.

42. Algunas reglas lingüísticas de lematización aplicadas en este proceso podrían ser: 1) llevar a singular los términos sancionados en plural, siempre y cuando tenga sentido hacer este cambio, o 2) preferir una forma masculina ante una femenina, si es pertinente. Otras reglas que se pueden aplicar, aunque no caen estrictamente en el terreno de la lematización, pero que mejoran la calidad de los datos al momento de hacer un análisis y que recomiendo considerar, son: 1) aplicar las reglas del uso de las mayúsculas y minúsculas, y 2) las reglas ortográficas y de acentuación de la lengua.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Como se puede ver en la figura 3, en este listado de los diez primeros candidatos a términos se registran: dos tipos de objetos: "indumentaria" y "pintura"; tres materiales: "barro", "algodón" y "madera"; y, por último, se registra una región: "Oaxaca". Además, hay cuatro unidades léxicas de las que, a simple vista, no resulta tan evidente que sean términos del dominio, éstas son: "Estado", "México", "enser" y "doméstico". Adelantándome a una explicación plausible, que se verá con mayor profundidad en el apartado siguiente, resulta que estas cuatro unidades léxicas forman los términos compuestos "Estado de México" y "enseres domésticos", este último en plural.

Por último, en la columna "Formas lematizadas" se pueden ver las unidades léxicas que fueron lematizadas, acompañadas de su frecuencia absoluta en el subcorpus, por ejemplo, para el caso de "madera" se mantuvieron 11, 724 formas singulares y se lematizaron ocho formas plurales ("maderas"). A partir de este hecho, se podría inferir que el material "madera" es mucho más común, en el discurso, en su "forma de palabra" singular y que la unidad de cita en un diccionario o vocablo debe ser "madera".

Búsqueda de términos pluriléxicos

Hay cuatro casos notables en los resultados del apartado anterior, éstos son "Estado", "México", "enseres" y "domésticos", de los que no resulta totalmente claro que sean términos monoléxicos y llama la atención, también, que dos de ellos se presenten siempre en plural, en el subcorpus analizado.

Como ya adelanté, estas cuatro unidades léxicas forman dos términos pluriléxicos; "Estado de México" y "enseres domésticos". La funcionalidad *Collocate* de AntConc permite sustentar esto a partir de las siguientes evidencias: "Estado", en el subcorpus, se combina regularmente con otras unidades léxicas en 29, 944 ocasiones, es decir, muy probablemente hay en el subcorpus cerca de 29, 000 ocurrencias de colocaciones (*collocate tokens*) que se pueden agrupar en 31 tipos. Estas colocaciones siempre se forman con la preposición: "de", seguida de los nombres de las 31 entidades federativas de México, es decir, formando: "Estado de México", "Estado de Zacatecas", "Estado de Yucatán", "Estado de Veracruz", y así sucesivamente. El caso de que "enseres" y "domésticos" formen un término pluriléxico compuesto es aún más contundente, ya que en el subcorpus: primero, ambas unidades léxicas siempre aparecen una seguida de la otra; segundo, invariablemente se encuentran en

https://doi.org/10.22201/iie.18703062e.2024.124.2861

plural, forman una frase nominal y, por tanto, establecen una relación de concordancia de número entre ellos; por último, la funcionalidad *Collocate* establece como único candidato para ser un colocado de "enseres" a "domésticos".

Por otra parte, ya se ha señalado que *madera* es uno de los candidatos a términos monoléxicos más frecuentes, obtenidos usando la funcionalidad *Word List* de AntConc, en combinación con una lista de filtrado y una serie de reglas de lematización; sin embargo, como es ampliamente conocido por todos, es común la existencia de términos pluriléxicos en los dominios científicos y técnicos, y más aún en esta área del saber. Por tanto, para inferir términos pluriléxicos que se construyen a partir de una base léxica, también se propone utilizar la funcionalidad *Collocate* definiendo una extensión máxima de la colocación.

De forma aplicada, he tomado como base el término "madera" y definido, en la funcionalidad *Collocate,* una extensión máxima del colocado de tres unidades léxicas, es decir, la funcionalidad de AntConc buscó agrupaciones pluriléxicas compuestas por dos y tres unidades léxicas que incluyan la base "madera" en primera posición. Los candidatos obtenidos, que muy probablemente sean términos del dominio, se enlistan a continuación:

madera de pino	madera de guayacán
madera de cedro	madera de naranjo
madera de campincerán	madera de tzompantle
madera de colorín	madera de balsa
madera de encino	madera de caoba

Los procesos y ejemplos ilustrados hasta ahora, aplicados a la obtención de candidatos a términos pluriléxicos, requieren el conocimiento *a priori* de, al menos, uno de los elementos léxicos que forman el término; sin embargo, otra posibilidad que ofrece AntConc es buscar grupos de expresiones comunes definiendo únicamente su longitud 'N', sin definir ninguna unidad léxica que compone el término. Para ello es necesario usar la función *Clusters/N-grams*, y definir un rango de longitud para las expresiones regulares. Al aplicar esta técnica al análisis del subcorpus se obtuvieron estos bigramas y trigramas que, muy probablemente, sean términos del dominio: "corteza de majagua", "medida de maíz", "papel de estraza", "tejido entrecruzado" o "traje huichol".

Como es natural, el resultado de este procedimiento enlista formas recurrentes en el discurso, no lexicalizadas, por ejemplo: "decidir los espacios",

https://doi.org/10.22201/iie.18703062e.2024.124.2861

"lucha contiene juegos", "ónix cristal de", "silbato armadillo forma" o "velas y alimentos". Inevitablemente, este procedimiento implica un trabajo arduo de revisión de candidatos, aunque también se pueden aplicar a los resultados una lista de filtrado o un cotejo respecto a un corpus de referencia que permita obtener resultados más precisos; proceso que será motivo de otra investigación y futura comunicación.

Inferencia de términos relacionados con la materialidad y técnicas de manufactura de "máscara"

En este último apartado se explica cómo inferir términos relacionados a un concepto utilizando la funcionalidad *Collocate* y, a partir de estos resultados, considerar léxico clave para crear definiciones terminológicas. Enseguida se ejemplifica cómo identificar términos relativos a la materialidad y técnicas de manufactura[43] de una "máscara", para ilustrar este punto.

En esta etapa del análisis exploratorio he restringido la experimentación a la observación de las técnicas y materiales relacionadas con un tipo de objeto ("máscara"), ya que, en el presente artículo, sólo pretendo mostrar las generalidades del uso de AntConc y no hacer un estudio terminológico elaborado de este término. Para ello he limitado la longitud del análisis a diez unidades léxicas a la derecha de cada ocurrencia de "máscara"; esto incluye siempre a las técnicas y los materiales que las describen, aunque no pierdo de vista que un análisis exhaustivo debe incluir también la procedencia de los objetos, el tema, la forma y otros datos contenidos en el subcorpus.

43. Para el ejemplo presentado se toman estos dos elementos de información comunes a la gran mayoría de modelos de metadatos usados para catalogar bienes de interés cultural. Si se quisiera saber más sobre la naturaleza de estas categorías descriptivas véase: Robin Thames, Robin Dorrell y Henry Lie, *Introduction to Object ID. Guidelines for Making Records That Describe Art, Antiques, and Antiquities* (Los Ángeles: Getty Information Institute, 1999); Visual Resources Association, "VRA Core 4.0 Element Description", 2007, https://www.loc.gov/standards/vracore/VRA_Core4_Element_Description.pdf (consultado por última vez el 11 de marzo de 2024); Murtha Baca y Patricia Harpring, "CDWA List of Categories and Definitions (Last Revised 9 April 2019)", https://getty.edu/research/publications/electronic_publications/cdwa/definitions.pdf (consultado por última vez el 11 de marzo de 2024); Panorea Gaitanou y Manolis Gergatsoulis. "Mapping VRA Core 4.0 to the CIDOC CRM Ontology", en *Proceedings of the First Workshop on Digital Information Management. March 30-31* (Corfu, Grecia: Ionian University, 2011), 26-38, entre otros.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

Frecuencia absoluta	Valor MI	Término
590	3.28810	tallado
525	3.15469	pintado
108	3.60487	esgrafiado
62	4.37354	confeccionado
59	3.00826	perforado
46	3.80964	labrado
33	3.02562	moldeado
27	2.99915	estucado
24	2.78858	aglutinado
19	3.62147	barnizado
17	3.13908	pulido
16	3.46665	curtido
15	2.52555	modelado
10	5.52555	laqueado
5	3.69547	aglutinado

4. Técnicas de manufactura relacionadas con "máscara".

Según los resultados ofrecidos por la funcionalidad *Collocate* de AntConc, en el subcorpus analizado, el término "máscara" tiene 8 950 ocurrencias de colocaciones (*collocate tokens*), de las cuales sólo hay 270 tipos (*collocate types*). Como se puede ver enseguida, en las figuras 4 y 5[44] se presentan, de forma resumida, los 15 resultados más frecuentes, según su frecuencia absoluta, tanto para materiales como para técnicas de manufactura implicadas en la confección de las máscaras descritas en el subcorpus; también se presenta el valor MI.

Como se puede ver en la figura 2, correspondiente a los materiales relacionados a "máscara", los más comunes con los que se fabrican éstas en México son madera, pintura, licra, piel, crin, pelo, cuero, papel, cuerno, barniz, cedro, barro, tela, cartón e ixtle; mientras que, en la figura 4 se enlistan las 15 técnicas de manufactura más comunes para este tipo de objeto, a saber: tallado,

44. Las figuras 4 y 5 son de elaboración propia, pero hay que aclarar que AntConc genera los resultados en ese mismo modelo tabular y lo visto en estas figuras es una representación fiel del análisis generado por la herramienta.

https://doi.org/10.22201/iie.18703062e.2024.124.2861

pintado, esgrafiado, confeccionado, perforado, labrado, moldeado, estucado, aglutinado, barnizado, pulido, curtido, modelado, laqueado y aglutinado. Con base en estos datos se puede anticipar que una definición de *máscara*, en el contexto nacional mexicano, que describa las manifestaciones de arte popular únicas de México, deberá considerar estos términos para explicar su materialidad y las técnicas de manufactura empleadas en su creación; además de señalar que las máscaras son "cubiertas para la totalidad o parte del rostro, generalmente con aberturas para los ojos y a veces la boca".[45]

Resumen y conclusiones

En este artículo se han explicado los pormenores del diseño de un corpus de catálogos de bienes artísticos y culturales de México, sincrónico, monolingüe, en formato electrónico, y con un claro enfoque al trabajo lexicográfico; recurso sin antecedente alguno.

Desde el punto de vista teórico, se ha argumentado sobre la importancia aplicativa de los corpus, sus tipos más generales y la noción de representatividad; y se ha concluido que esta última es modelable, si se cuenta con información de la población estadística que se quiere representar; si no fuera el caso, se puede tratar al corpus como una muestra primordial y, *grosso modo*, integrar el mayor número de materiales disponibles. Asimismo, se ha señalado que la representatividad puede verse afectada en favor del uso del corpus en tareas específicas, tales como "buscar representar el mayor número de términos posibles", aunque esto implique un sesgo en la selección documental.

Al aceptar el argumento de que las fuentes primarias "no textuales" de la historia del arte y la arqueología (restos óseos, fotografías, pinturas murales, monumentos…) se describen verbalmente en catálogos, así como la importancia capital de este género textual para el dominio, el sesgo de este corpus, criticable desde la teoría planteada, justifica la selección de los documentos que se incluirían en el recurso y los objetivos de éste.

También se hizo una brevísima revisión de sistemas de gestión de corpus y se ha argumentado sobre las razones operativas por las que AntConc es una herramienta muy conveniente para esta propuesta, a saber, el gestor cuenta

45. The Getty Research Institute y Centro de Documentación de Bienes Patrimoniales, *Tesauro de Arte & Arquitectura*, https://www.aatespanol.cl/ (consultado por última vez el 11 de marzo de 2024).

https://doi.org/10.22201/iie.18703062e.2024.124.2861

con múltiples funciones para lexicografía y otras funciones avanzadas, es multiplataforma (Mac OS, Windows y Linux), no requiere instalarse, es fácil de manejar, independientemente del perfil de usuario; es idóneo para el manejo de corpus pequeños y medianos y es un *software* libre.

En términos aplicados, un corpus como el propuesto permitiría identificar, mediante el uso de herramientas y técnicas de análisis textual, términos monoléxicos y pluriléxicos, así como documentar sus contextos de uso en catálogos de museos e instituciones nacionales; esto, para ir creando un fichero terminológico que sirva como base empírica para la redacción de definiciones de términos del dominio.

He ilustrado un caso puntual de procesamiento de la información de los documentos con la intención de mantener "sentidos completos", y cómo opté por conservar los catálogos en un formato ideal para su preservación, y en archivos independientes, para que se puedan modelar búsquedas dependiendo de las necesidades de los usuarios potenciales del corpus. En resumen, la preparación de los catálogos implica estos pasos: 1) seleccionar la documentación; 2) organizar los elementos de información en un orden de prelación; 3) eliminar datos sensibles y los que podrían generar ruido, y 4) transformar los datos de los catálogos a un formato de texto plano (.txt) para preservarlos efectivamente. También se han mostrado algunos resultados de la aplicación de esta metodología y de las herramientas de análisis textual a un subcorpus (obtención de candidatos a términos y de léxico que podría servir para definirlos desde la perspectiva de su materialidad y manufactura).

Por último, hay que destacar que, aunque esta propuesta es de tipo aplicado, de ella se derivan una metodología de procesamiento de textos y un esbozo de posibilidades de aplicación de herramientas de análisis textual de uso terminográfico que también pueden emplearse en la investigación y enseñanza de la historia del arte y arqueología. ♣