FUNCIONALIDADES DEL SISTEMA DE GESTIÓN DE CORPUS, GECO, PARA LA ELABORACIÓN DE DICCIONARIOS DE ESPECIALIDAD

Functionalities of the Corpus Manager System, GECO, for the Preparation of Specialized Dictionaries

Gerardo Sierra[1]

RESUMEN

Es innegable el uso de herramientas computacionales para la labor lexicográfica en la creación de diccionarios de especialidad. El proceso va desde el diseño y constitución del corpus de especialidad, pasando por la extracción de los términos, hasta la extracción automática de definiciones, lo que constituye un trabajo de la lexicografía computacional. Este artículo esboza, en primer lugar, los elementos por considerar en la constitución del corpus de especialidad. Posteriormente se describe el sistema GECO, un gestor para subir el corpus, así como las aplicaciones de este sistema para la construcción de diccionarios, tales como el extractor automático de términos, el visualizador de concordancias y el extractor de contextos definitorios Con ello, se muestra una solución efectiva a los problemas en el proceso de elección de términos y la extracción de definiciones desde un corpus de gran envergadura, lo cual reduce significativamente el tiempo de procesamiento de los datos obtenidos con el fin de agilizar la creación de un diccionario especializado.

Palabras clave: Lexicografía computacional, extracción de información, procesamiento de lenguaje natural.

ABSTRACT

The use of computational tools for lexicographical work in the creation of specialty dictionaries is undeniable. The process starts from the design and constitution of the specialty corpus, through the extraction of terms, to the automated extraction of definitions, which is a work of computational lexicography. This paper sketches in first place the elements to be considered in specialty corpus constitution. Subsequently the GECO system is described, a manager to upload the corpus, as well as the system applications for dictionaries construction such as the automatic term extractor, the concordances viewer and the definitional contexts extractor. This approach shows an effective solution to the problems in the process of choosing terms and extracting definitions from a large corpus, which significantly reduces the processing time of the data obtained in order to improve the creation of a specialized dictionary.

Key Words: Computational Lexicography, information extraction, natural language processing.

[1] Universidad Nacional Autónoma de México. Investigador titular, Grupo de Ingeniería Lingüística. México.

Correo electrónico: gsierram@iingen.unam.mx

Recepción: 5-5-2019. *Aceptación: 1-8-2019.*

1. Introducción

Los textos especializados son una fuente necesaria para la elaboración de diccionarios de especialidad. En estos textos se encuentran los términos del área correspondiente, así como información relevante sobre estos para obtener los elementos necesarios para su definición. Como ejemplo, se muestra el siguiente fragmento de un texto obtenido del Corpus Técnico del IULA, a través de BwanaNet[2] (cf. Bach, Saurí, Vivaldi y Cabré, 1997):

> Los compuestos **que** no derivan de la adormidera, **pero que** ejercen efectos directos uniéndose a los receptores específicos para opiáceos **se denominan** opioides. Desde un punto de vista práctico, los opioides **se definen como** compuestos de acción directa, **cuyos** efectos se ven antagonizados estereoespecíficamente por la naloxona. **Las tres clases principales** de opioides endógenos en los mamíferos **son** las encefalinas, las endorfinas y la dinorfina (s. p.).

En este texto es posible identificar los términos del área, pero también se observan ciertos patrones que permiten extraer las definiciones brindadas en el texto, así como algunas relaciones semánticas entre los términos. En efecto, en negritas aparecen los relativos ‘que’ y ‘cuyo’, así como las predicaciones ‘se denominan’, ‘se definen como’ y ‘las tres clases principales (…) son’, los cuales constituyen patrones regulares que permiten identificar los términos, las relaciones léxicas y las definiciones, con lo que se construye un diagrama como el siguiente (Figura 1):

[2] Disponible en http://bwananet.iula.upf.edu/

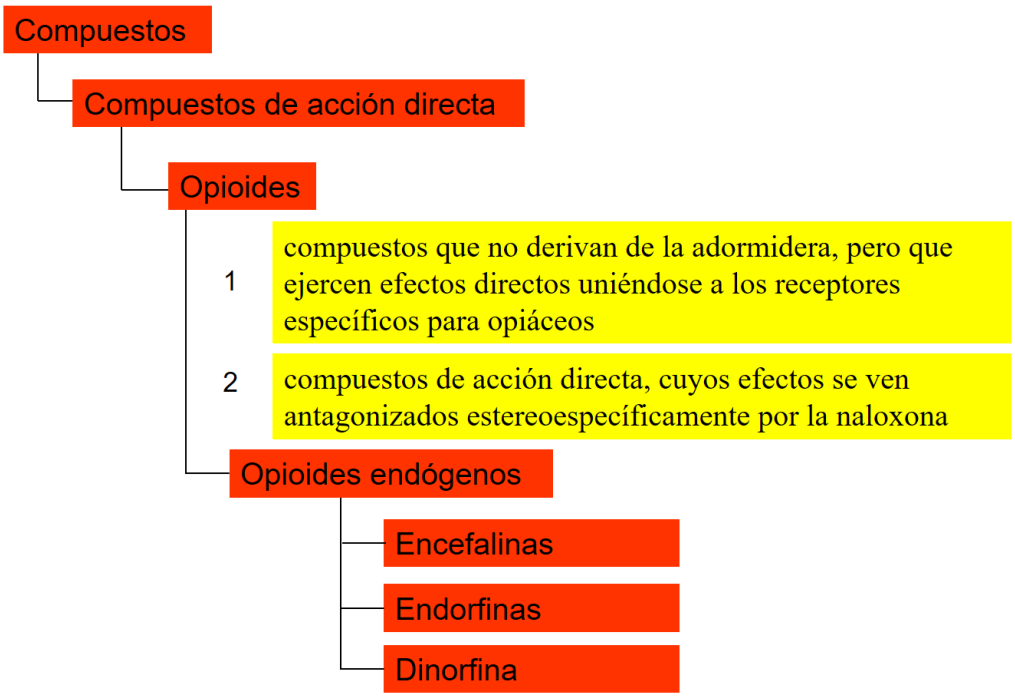

Figura 1. Ejemplificación de términos y definiciones extraídos de un texto

Fuente: Elaboración propia.

Mediante la identificación de estos patrones regulares es posible extraer automáticamente de los textos de especialidad los elementos necesarios para la construcción de diccionarios. En este sentido, a continuación, se presentan algunas herramientas informáticas que se utilizan, en general, para la elaboración de diccionarios, en concreto las herramientas desarrolladas por el Grupo de Ingeniería Lingüística de la UNAM (cf. Sierra, Lázaro y Bel-Enguix, 2016), disponibles en línea y de libre acceso. En primera instancia se describe la importancia de los corpus lingüísticos, las características que deben cumplir y se enfatiza el Gestor de Corpus GECO. Luego se describe el proceso de extracción de términos y la herramienta Termext. Con ello, se describe la extracción de definiciones y el Extractor de Contextos Definitorios ECODE.

2. Corpus lingüísticos

Resulta innegable el papel de los corpus lingüísticos para la elaboración de diccionarios (cf. Rojo, 2009). El diseño y la creación de un corpus lingüístico que busca contar con material original desde su inicio conlleva muchas más tareas de las que se podría imaginar y es, de hecho, la parte que más tiempo lleva en una investigación lingüística. Es de notar que un corpus lingüístico es solo una muestra de la lengua y no un reflejo de la totalidad de ella, por lo que, al diseñar un corpus se debe intentar que sea representativo (cf. Kabatek, 2013), pero que al mismo tiempo tenga una amplia cobertura y no se incline hacia algún registro, es decir, que sea balanceado (cf. Torruella, 2017). Sin embargo, existe una inmensa variedad de registros, cambios, fenómenos o grupos sociales existentes en una comunidad, entre ellos:

- **Localidad geográfica.** La representatividad en un corpus puede verse como una convergencia de textos provenientes de diversas localidades geográficas, dentro de un mismo territorio. Así, en un corpus del habla de Costa Rica, para ser representativo se deben tomar en cuenta las muestras de cada una de las siete provincias, y no solo de las más importantes económicamente o las más densamente pobladas, por ejemplo.
- **Información personal.** Los datos personales del informante pueden ser decisivos para los estudios léxicos y pragmáticos, por lo que habrá que diferenciar, entre otros datos, el género, la edad o grupo etario, el estrato sociocultural, el nivel de estudios y la preferencia sexual.
- **Tópico.** La distinción del tópico, entendido como el tema o asunto del que se habla, será siempre subjetiva pero necesaria, por lo que conviene apoyarse en expertos en el tema en torno al cual girará dicho corpus, pues los especialistas son quienes conocen con mayor precisión la tipología del área y las fuentes más importantes para la extracción de documentos.

- **Tipo de texto.** Un texto oral y un texto escrito estarán casi siempre destinados a estudios muy distintos y, por tanto, los rasgos de cada uno deben estar bien documentados y diferenciados.
- **Fuente.** La fuente explícita del texto es un dato que siempre debe existir y proporcionarse de manera obligatoria en un corpus. El soporte de un texto puede ser un libro, una página de Internet, un manuscrito, etc. Por otra parte, en la obtención de un texto oral se pueden nombrar los programas de radio, las entrevistas o la televisión.
- **Tiempo.** Ya sea sincrónico o diacrónico. El corpus lingüístico debe tener una referencia temporal bien definida con el fin de que los estudios hechos con base en él sean confiables y cumplan con su objetivo, es decir, para que haya coherencia entre el estudio que se va a realizar y el objeto de estudio.

La variedad es un criterio de mucha importancia para los corpus, porque de esta depende que el conjunto de textos refleje con claridad el universo de rasgos distintivos de una lengua, en un territorio y tiempo determinados. La variedad persigue la idea de que hay que documentar cada una de las variantes de la lengua que se esté estudiando para llegar a la meta de representatividad descrita.

El equilibrio tiene que ver con los límites de lo que se toma en consideración para un corpus y lo que no. Unido estrechamente con el tamaño del corpus, el equilibrio intenta hacer de la colección de textos una estructura homogénea que no se vea afectada o influenciada por alguno de sus subtemas o tópicos. Equilibrio significa contar con una muestra representativa y variada de alguna lengua siempre y cuando la distribución de los textos sea, en el mejor de los casos, equitativa tipológicamente o por las divisiones hechas anteriormente por el compilador y sus colegas.

Un ejemplo de corpus de especialidad, razonablemente variado, representativo y equilibrado desarrollado por el Grupo de Ingeniería Lingüística se tiene en el Corpus de las Sexualidades en México, CSMX,[3] el cual fue diseñado para estudiar y extraer la terminología del español de México (cf. Sierra, Medina y Lázaro, 2012). La distribución del corpus es piramidal, dividido en nueve áreas,

[3] Disponible en http://www.corpus.unam.mx/csmx.

y estas, a su vez, en cuatro niveles. La división del gran tema de la sexualidad en áreas y niveles se realizó con el fin de tener un alcance óptimo en lo que constituye esta área biológico-social-cultural, y que además se apegue a la realidad lingüística completa del área de especialidad, pero sin olvidar el léxico coloquial.

Así, la variedad cubre criterios diatécnicos, esto es, la variación léxica de acuerdo con hablantes de la misma área de especialidad o sociolecto, que considera las áreas delimitadas por los especialistas del área. Para cumplir con la representatividad, se definió que la terminología describiera el español mexicano de acuerdo con una división en criterios distráticos. Esta diferenciación por grupo social desembocó en una clasificación por el origen de los documentos: academia, foros, asociaciones, etc., cada uno de los cuales delimita su contenido por el tipo de hablantes que lo conforman. Para ello, se compilaron desde artículos en Google académico, artículos en revistas, foros, periódicos, chats, etc. También, se transcribieron algunas entrevistas hechas a estudiantes de Pedagogía y Antropología. De esta manera, se tenían textos con registros formales y coloquiales, pasando por registros estándares y vulgares, como se aprecia en la Tabla 1.

Tabla 1. Conformación del Corpus de las Sexualidades de México

Base estadística del CSMX (número de palabras)						
Áreas	**Científico**	**Informativo**	**Foros**	**Ficción**	**Total**	**%**
Fundamentos	90,546	23,698	62,088		176,332	11.54%
Respuesta	5,112	11,410	24,679		41,201	2.70%
Comportamiento	17,221	10,622	34,476		61,902	4.05%
Identidad	59,153	6,404	34,059		99,616	6.52%
ETS	32,901	12,055	24,661		69,617	4.56%
Parafilias	14,224	21,320	27,908		63,452	4.15%
Atracción	62,996	31,505	43,759		138,260	9.05%
Educación	66,402	15,523	28,298		110,223	7.21%
Literatura				767,300	767,300	50%
Número de palabras	348,555	132,537	279,928	767,300	1,527,903	100.00%
%	22.81%	8.67%	18.32%	50.22%	100.03%	100.00%

Fuente: Elaboración propia a partir del Corpus de las Sexualidades en México.

En otras palabras, la variedad está íntimamente ligada a la representatividad, ya que la primera condiciona a la segunda cuando la diversidad de textos que se buscan se ligan, por un lado, al propósito del estudio al que va dirigido el corpus y, por otro, a los estratos que serán tomados en cuenta para la extracción de dichos textos.

3. Geco

GECO[4] es un sistema de gestión colaborativa de corpus basado en web, desarrollado por el Grupo de Ingeniería Lingüística de la UNAM (cf. Sierra, Solórzano y Curiel, 2017), que permite a los usuarios subir colecciones de documentos y volverlos corpus digitales, clasificados por los metadatos, y los documentos anotados con etiquetas de partes de la oración, mediante Freeling (Padró y Stanilovsky, 2012). Con ello, el corpus es procesado con diversas aplicaciones implementadas como módulos integrados a la infraestructura de GECO.

El primer aspecto de GECO es que tiene la finalidad de ser un repositorio central de documentos para construir corpus. Los usuarios pueden agrupar documentos en carpetas y anotar sus metadatos, estos últimos posteriormente permitirán la recuperación y filtrado de los documentos. Para construir un corpus, uno de los aspectos flexibles del sistema es que, dentro de un marco de colaboración, permite a los usuarios escoger archivos individualmente de varias carpetas y agruparlos en un proyecto, y asignar un nombre a dicha selección: el nombre del corpus. Para el sistema, un proyecto es una colección de documentos, incluso contenidos en diferentes carpetas, unificados con un mismo nombre y descripción.

[4] Disponible en http://www.corpus.unam.mx/geco.

Cuando se crea un proyecto, el sistema permite al propietario crear un portal web para dicho proyecto, en el cual los internautas pueden consultar información acerca del corpus, tal como los participantes, agradecimientos, publicaciones relacionadas, entre otros.

La idea de GECO es ser útil a la comunidad, ofreciendo una *suite* de herramientas de análisis para un usuario que no sea experto en computadoras. Su diseño modular fue pensado para integrarse con diversas aplicaciones de corpus, las cuales fueron diseñadas para aprovechar el contenido de los proyectos de GECO. Las aplicaciones son calculadas a partir de los proyectos registrados en el catálogo de GECO. La aplicación hace una solicitud a la API de GECO, la cual retorna una lista de proyectos disponibles y a los que el usuario tenga permiso de acceso según las políticas de seguridad manejadas por GECO, a fin de que seleccione qué corpus utilizará.

Las tres aplicaciones que se describen a continuación son las que resultan de interés para la labor lexicográfica: el extractor de términos, el extractor de concordancias y el extractor de contextos definitorios.

3.1 Extracción de términos

Una herramienta ampliamente utilizada en la lexicografía de especialidad y terminología es el extractor de términos, ya que justo a partir de los textos de especialidad proporcionados por los corpus lingüísticos es posible obtener la terminología respectiva, que corresponderá a la entrada del diccionario (cf. Conrado, Pardo y Rezende, 2014).

Con los medios de comunicación actuales y la llamada globalización existe una marcada tendencia a la estandarización de las cosas, incluyendo la terminología. Si bien los términos de un área tienden a ser los mismos en el ámbito internacional, existen algunas variaciones. Asimismo, hay diferencias entre las palabras de uso no especializado y las que se refieren a los conceptos de especialidad; dependiendo de la región geográfica o el estrato social van cambiando, de acuerdo con el usuario y manifiestan las características de este (cf. Lara, 2016). Por tanto, en un diccionario de

especialidad es necesario incluir tanto los términos del área de especialidad, como las palabras de uso no especializado. Una manera de encontrar los términos del área de especialidad utilizados por las autoridades del tema, así como aquellas palabras utilizadas por la población general para referirse a los mismos términos, es la creación de un corpus de documentos con dicha información.

El método clásico consiste en comparar la distribución de palabras de un corpus de propósito general contra el corpus de especialidad. El método consiste en seleccionar aquellas palabras que tienen la más alta frecuencia del corpus de especialidad que no se encuentran en el corpus de propósito general. Los candidatos a término se presentan en una distribución uniforme de términos y se ordenan según su frecuencia de uso. Un problema común resulta con los términos que aparecen con muy baja frecuencia, ya que acaban discriminados.

En contraste, se han creado métodos que toman en cuenta las reglas lingüísticas de la formación de términos, tales como su morfología y su categoría sintáctica. El estado del arte en el área ha mostrado que los métodos híbridos obtienen mejores resultados, esto es, los que se basan en métodos estadísticos y en conocimiento lingüístico.

Gracias a la colaboración de la Universidad de Manchester y de la Universidad de Montreal, se desarrolló en el Grupo de Ingeniería Lingüística un primer sistema extractor de términos para el español, Termext, el cual consiste en un método híbrido que utiliza tanto conocimiento lingüístico como métodos estocásticos (Barrón, Sierra, Drouin y Ananiadou, 2009). Termext realiza algunas adaptaciones lingüísticas y funcionales al algoritmo C-value/NC-value que funciona como la base del grupo TerMine para reconocer términos candidatos multipalabras a partir de documentos especializados en inglés (Frantzi, Ananiadou y Mima, 2000). En este sentido, son dos las principales aportaciones del Termext: extrae términos en español y los candidatos pueden ser monopalabra o multipalabras.

A grandes rasgos, el método consta de dos partes: la obtención de C-Value y la de NC-Value. La primera utiliza las etiquetas de partes de la oración y de lemas para aplicar un filtro lingüístico que identifica las estructuras que pueden formar un término en español y calcula la probabilidad de que

dicha estructura sea un término en función de la frecuencia de la estructura, la frecuencia de la estructura en estructuras más grandes, el número de ocurrencias de las estructuras más grandes anteriores y la longitud de la estructura. La segunda considera la relevancia del contexto en el que se encuentran los términos para identificar qué tan representativo es el término o no. Los términos con valores más altos de NC-Value son los más importantes en el documento, en tanto que los de menor valor no son términos tan representativos (Figura 2).

Resultados del proyecto - **"CSMX completo"**

Mostrar 10 **términos por página** **Buscar:** relación

#	Término	Frecuencia	NC-Value
29	relación sexual	562	903.360156268
72	relación	609	509.711425026
1086	relación sexual coital	23	49.5581785594
1153	relación de género	22	46.339698581
1537	relación social	25	33.6890831662
1842	relación de pareja	27	27.6319533715
2000	relación de poder	17	25.2816520437
2036	relación amoroso	21	24.8518213494
2772	relación coital	16	17.1123725055
3197	frecuencia de relación sexual	6	14.7655749129

Mostrando 1 al 10 de 16 términos
(filtrados de 4,999 términos totales)

Ant Sig

Figura 2. Ejemplo de resultados del extractor terminológico termext para el corpus de las sexualidades de México

Fuente: Elaboración propia a partir del Corpus de las Sexualidades en México.

Termext recibe los textos contenidos en un proyecto registrado en GECO y proporciona como resultado una lista de términos con una alta probabilidad de pertenecer al área de conocimiento del proyecto, ordenados por puntaje, según el valor de NC-Value, aunque también presenta el lugar que

ocupa el término en la lista y la frecuencia de este. También es posible buscar los resultados que cumplen con una secuencia dada. Como se observa de la Figura 2, para el Corpus de las Sexualidades en México aparecen los términos que contienen la palabra *relación*.

La aplicación de Termext permite con un solo clic sobre el término ir a la siguiente aplicación de GECO: Concordancias.

3.2 Concordancias

Una vez que se identifican los términos de un área de especialidad, en segundo lugar se realiza un estudio de estos para encontrar su significado. Para esto, una herramienta tradicional es el uso de concordancias (cf. Faber, Moreno y Pérez, 1999).

Las concordancias son listas del contexto en el cual se encuentra un término extraídas de un conjunto de documentos, a partir de una petición por un usuario. Esto se conoce como Keyword in context o KWIC, que se presenta en tres columnas: la del centro muestra la petición del usuario, en tanto las columnas de los lados muestran el contexto, esto es, las palabras que aparecen a la izquierda y a la derecha. Las concordancias son el punto de partida para la elaboración de diccionarios de especialidad. Tan solo tómese en cuenta el análisis de concordancias a partir del Corpus del Español Mexicano Contemporáneo que se llevó a cabo para constituir el Diccionario del Español de México (cf. Lara, Ham Chande y García Hidalgo, 1979).

La aplicación de Concordancias en GECO aprovecha que los documentos han sido previamente indexados y anotados automáticamente con partes de la oración. Con ello, el usuario puede ejecutar consultas eficientes, ya sea búsquedas de palabras, de lemas o de categorías gramaticales; búsquedas de proximidad, especificando la distancia de una cadena con respecto a otra; y con filtro de los documentos basados en sus valores según los metadatos previamente definidos (Figura 3).

Izquierda	Petición	Derecha
los sexos , a_la_vez_que se constituye en una forma primaria	de relaciones significantes	de poder . El género muestra el carácter de tens
el dominio masculino definido por la exclusividad y multiplicidad	de relaciones heterosexuales	; la visión desintegrada de el cuerpo femenino c
básico para su consecución , lo cual cuestiona los modelos	de relaciones sociales	e institucionales que han sido construidas para
de una pareja (otras variables intermedias serían la frecuencia	de relaciones sexuales	, el aborto y la lactancia) , el cual
sociedad donde estes . La sociedad , es un cumulo	de relaciones humanas	donde todos depositan sus pensamientos para
. 2 . - La sociedad , es un cumulo	de relaciones humanas	donde todos depositan sus pensamientos para
cada vez más la asume . Otra cosa sería hablar	de relaciones alternativas	, basadas en el respeto a sí mismo , a
pensamiento materno , ya_que este pensamiento es resultado	de relaciones sociales	, es decir , se construye a_través_de un proces
el mismo . Se les debe insistir en abstener se	de relaciones sexuales	durante los tratamientos . La pareja sexual debe
Sin_embargo , las ITS pueden transmitir se en cualquier forma	de relación sexual	, así que es importante que todos los miembros
, desde el siglo_XIII , Occidente ha remitido sus deseos	de relaciones interpersonales	sexuadas (o sublimadas) " , como lugar utópic
) Vivir el resto de tu vida sin ningún tipo	de relación amorosa	. Tus padres te seguirán queriendo , tendrás am

Figura 3. Ejemplo de concordancias con partes de la oración para el Corpus de las Sexualidades de México

Fuente: Elaboración propia a partir del Corpus de las Sexualidades en México.

En la Figura 3 se observan los resultados para una búsqueda en el Corpus de las Sexualidades en México. La búsqueda es la siguiente:

de [relación] <AQ*>

Los comodines * y ? buscarán cualquier subcadena y cualquier caracter, respectivamente. Entre corchetes angulares se busca una etiqueta de partes de la oración conforme al etiquetado definido por el grupo EAGLES.[5] En la consulta del ejemplo, se busca la categoría de adjetivo calificativo, independientemente de sus atributos (grado, género, número o función). El corchete permite buscar todas las formas del lema, en este caso para el lema ‘relación’ se encuentra ‘relaciones’ y ‘relación’. Además, es posible realizar búsquedas de proximidad (no se muestra en el ejemplo anterior), las cuales especifican la distancia de una cadena con respecto a otra. Para ello, se escribe entre llaves un número que corresponde a la distancia deseada entre las dos cadenas.

Los colores de las palabras en la Figura 3 aparecen cuando el usuario solicita ver las etiquetas de partes de la oración. Tómese en cuenta que, dado que GECO utiliza Freeling para etiquetar automáticamente, puede haber errores.

[5] Disponible en http://www.lsi.upc.es/~nlp/tools/parole-sp.html.

Debido a que los metadatos capturados en GECO quedan registrados, es posible filtrar documentos basados en sus valores; esto permite efectivamente crear subcorpus en determinado momento. Por ejemplo, para el Corpus de las Sexualidades en México puede crearse un subcorpus exclusivo para ficción, o bien otro para documentos de las áreas de comportamiento, atracción y parafilias. Para esta clase de filtrado la interfaz presenta selectores de pares campo-valor para restringir el dominio de búsqueda. Asimismo, GECO permite observar las gráficas de la distribución de los documentos obtenidos en una consulta, según los metadatos que se hayan seleccionado, como se muestra en la Figura 4.

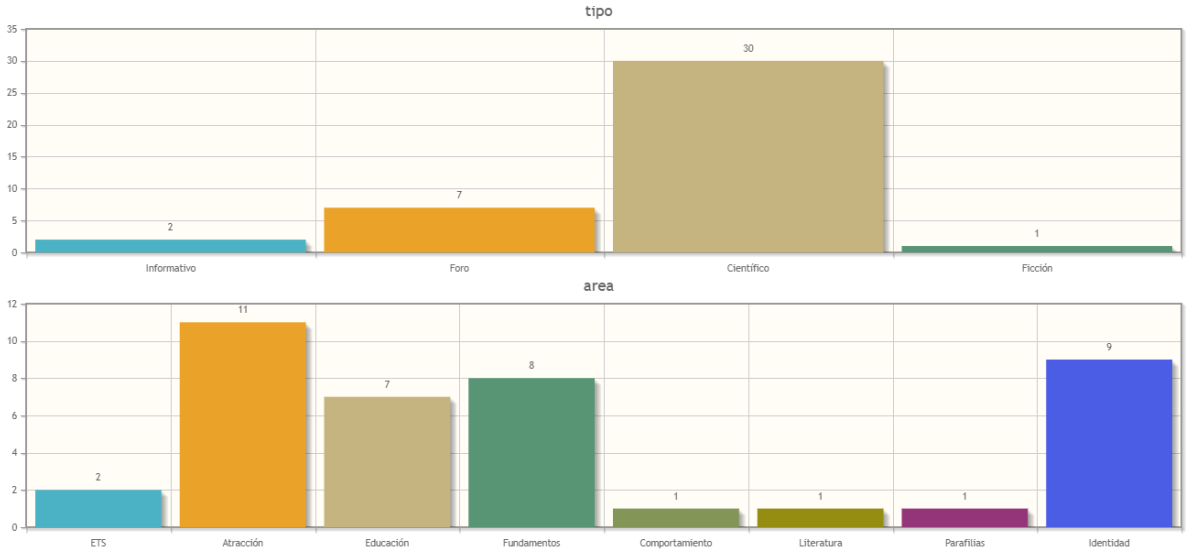

Figura 4. Gráficas para una búsqueda del Corpus de las Sexualidades de México

Fuente: Elaboración propia a partir del Corpus de las Sexualidades en México.

3.3 Extracción de contextos definitorios

Actualmente existe un creciente interés por el desarrollo de sistemas para la identificación automática de información sobre términos que sean útiles para describir su significado. Si bien la herramienta de concordancias es comúnmente usada en este sentido, pueden aprovecharse los textos de especialidad para extraer aquellos fragmentos en que los autores proporcionan las definiciones de los términos que introducen. En efecto, cuando el autor de un texto especializado define un término

lo hace mediante ciertas estructuras en las cuales se suelen utilizar una serie de patrones léxicos y metalingüísticos que son reconocidos automáticamente (cf. Pearson, 1998).

Estas estructuras que pueden ser extraídas han sido denominadas contextos definitorios (cf. Sierra, 2009). Un contexto definitorio es un fragmento textual de un documento especializado donde se aporta la definición de un término. Según el tipo de definición que se proporcione, se proporcionan además datos sobre las características y atributos del término, así como funciones, partes o bien relaciones de este con otros términos. Con ello, el uso de contextos definitorios se extiende no solo a la elaboración de diccionarios de especialidad (cf. Corrales, Miranda, Casasola, Leoni y Hernández, 2018), sino a la extracción de relaciones de hiponimia/hiperonimia (Aguilar y Acosta, 2016), entre otras.

Como desarrollo del GIL en el campo de la ingeniería lingüística, se crea Ecode, un sistema extractor de contextos definitorios para el español, el cual trabaja a partir de corpus lingüísticos, clasificándolos en tres tipos según su definición (analítica, funcional y extensional) e identificando sus dos principales elementos constitutivos: término y definición (cf. Alarcón, Sierra y Bach, 2008).

El principio para extraer contextos definitorios de manera automática lo constituye la identificación de los patrones verbales definitorios (PVD), que son las construcciones sintácticas verbales que unen un término con su definición. Entre estos verbos se encuentran principalmente los metalingüísticos (por ejemplo, ‘definir’, ‘entender’ o ‘denominar’), aunque también otros empleados en diferentes situaciones comunicativas –no solo definitorias–, como los verbos ‘ser’ y ‘considerar’. Aunado a estos verbos, en el PVD se toman en consideración algunas partículas gramaticales como el pronombre impersonal ‘se’ en posición proclítica o enclítica en relación con el verbo definitorio; las preposiciones ‘a’ o ‘por’, y el adverbio ‘como’. De esta manera, ejemplos de PVD son los siguientes: ‘entendemos por’, ‘denominarse a’ o ‘se define como’.

Una vez que se extraen las oraciones a partir de los PVD, se eliminan aquellos contextos que no son relevantes, según unas reglas de filtrado y, posteriormente se identifican los elementos constitutivos de los contextos definitorios. El filtro de contextos no relevantes se basa en una serie de

reglas lingüísticas y contextuales para determinar los casos en los que es probable que un patrón verbal no esté introduciendo información definitoria. Para la identificación de los elementos constitutivos, esto es, el término y la definición, se utiliza un árbol de decisión que recurre a la gramática de patrones verbales. El árbol de decisión, a través de inferencias lógicas, asocia una serie de patrones contextuales para cada verbo definitorio, de forma que dichos patrones indican las posiciones en que puede aparecer el término y la definición.

Los contextos definitorios que resultan después del filtrado de excepciones de contextos no relevantes, y con la identificación de los elementos constitutivos, se agrupan primero por el término y después por el tipo de definición.

4. Conclusiones

La elaboración de los diccionarios precisa de gran dedicación, tiempo y esfuerzo. Si bien la informática, en general, ha facilitado la labor del lexicógrafo, aún existe la necesidad de herramientas computacionales especializadas para esta labor. El avance tecnológico en el desarrollo de herramientas para facilitar el trabajo lexicográfico provee de corpus lingüísticos especializados donde se almacena digitalmente una gran cantidad de documentos técnicos. Una vez que se cuenta con estos recursos, existen herramientas para la extracción automática de términos, la visualización de concordancias para entender el significado de los términos y la extracción de contextos definitorios.

Con la integración de las herramientas mostradas en GECO, se tiene un recurso sumamente valioso en la práctica lexicográfica. La extracción semiautomática de diversas terminologías y la creación de diccionarios pueden ser una realidad que incluya un procesamiento más veloz y fino al combinar las técnicas de la lexicografía clásica con la inclusión de la lexicografía computacional y la extracción automática de definiciones.

Cabe resaltar que estas herramientas pueden integrar al conocimiento del lexicógrafo información actualizada día tras día con la revisión de los corpus y el procesamiento de las

aplicaciones. De esta manera, conscientes sobre el cambio semántico de las palabras, se puede llevar a cabo un análisis más exacto y minucioso de los nuevos términos que se van acuñando y de sus significados inherentes.

Referencias bibliográficas

Aguilar, C. y Acosta, O. (2016). Design of a Extraction System for Definitional Contexts from Biomedical Corpora. *BAI@IJCAI*.

Alarcón, R., Sierra, G. y Bach, C. (2008). ECODE: a pattern based approach for definitional knowledge extraction. *Proceedings of the XIII EURALEX International Congress.* Barcelona, España, 923-928.

Bach, C., Saurí, R., Vivaldi, J. y Cabré, M. T. (1997). *El corpus técnico de l'IULA: descricpió.* Informe 17. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona, España.

Barrón, A., Sierra, G., Drouin, P. y Ananiadou, S. (2009). An Improved Automatic Term Recognition Method for Spanish. *Lecture Notes in Computer Science*, *5449*, 125-136.

Conrado, M., Pardo, T. y Rezende, S. (2014). The main challenge of semi-automatic term extraction methods. *Proceedings of the 11st International Workshop on Natural Language Processing and Cognitive Science* – NLPCS. Venecia, 27-29.

Corrales Montero, S., Miranda Hernández, K., Casasola Murillo, E., Leoni de León, J. A. y Hernández Delgado, M. (2018). Análisis de texto para la identificación automática de marcadores lingüísticos definicionales en recetas de gastronomía de Costa Rica. *Kañina*, *Revista de Artes y Letras de la Universidad de Costa Rica*, *XLII* (3), 65-78.

Faber, P., Moreno, A. y Pérez, C. (1999). Lexicografía computacional y lexicografía de corpus. *Revista Española de Lingüística Aplicada*, *1*, 175-214.

Frantzi, K., Ananiadou, S. y Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, *3* (2), 115-130.

Kabatek, J. (2013). ¿Es posible una lingüística histórica basada en un corpus representativo? *Iberoromania*, *77*, 8-28.

Lara, L. F. (2016). *Teoría semántica y método lexicográfico*. México: El Colegio de México.

Lara, L. F., Ham Chande, R. y García Hidalgo, I. (1979). *Investigaciones lingüísticas en lexicografía*. México: El Colegio de México.

Padró, L. y Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. En *Language Resources and Evaluation Conference (LREC'2012)*, 2473-2479.

Pearson, J. (1998). *Terms in Context.* Philadelphia, EE. UU.: John Benjamins.

Rojo, G. (2009). Sobre la construcción de diccionarios basados en corpus. *Revista Tradumática, Traducción y Tecnologías de la Información y la Comunicación*, *7*, 1-7.

Sierra, G. (2009). Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática*, *1*(2), 13-37.

Sierra, G., Lázaro, J. y Bel Enguix, G. (2016). LEXIK: An integrated system for specialized terminology. *Lecture Notes in Artificial Intelligence*, *10061*, 79-91.

Sierra, G., Medina, A. y Lázaro, J. (2012). Terminótica y sexualidad: un proyecto integral. En *Actas XII Simposio Iberoamericano de Terminología*. Red Iberoamericana de Terminología RITerm, Buenos Aires, Argentina, 84-102.

Sierra, G., Solórzano, J. y Curiel, A. (2017). GECO, un gestor de corpus colaborativo basado en web. *Linguamatica*, *9* (2), 57-72.

Torruella, J. (2017). *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística*. Frankfurt am Main: Peter Lang Ed.