Bases de datos relacionales, una herramienta digital para la gestión del análisis semántico

Ernesto Wong García

Bases de datos relacionales, una herramienta digital para la gestión del análisis semántico

Relational Databases, a Digital Tool for Managing Semantic Analysis

Autor: Ernesto Wong García
0000-0002-7548-1342 (ORCID)
ewong@flex.uh.cu, ernestowg@gmail.com
Facultad de Lenguas Extranjeras, Universidad de La Habana
Avenida 19 de Mayo #14, esq. Amézaga
Plaza de La Revolución, La Habana, Cuba

Resumen:

Entre las muchas herramientas digitales disponibles para la investigación lingüística, se hace notar la ausencia de herramientas que permitan gestionar el análisis semántico y lingüístico en general, manteniendo los datos organizados, estructurados y recuperables, y brindando la flexibilidad para elaborar aparatos analíticos propios. En esta nota, consideramos las bondades que, en este sentido, ofrecen las bases de datos relacionales y referimos la experiencia de un proyecto lexicográfico. Reseñamos las principales características y el funcionamiento de las bases de datos relacionales, describimos el diseño de la utilizada en dicho proyecto y presentamos los beneficios que nos ha reportado el uso de esta herramienta.

Palabras clave:

bases de datos relacionales, semántica, análisis lingüístico, lexicografía, herramientas digitales

Abstract:

Among the many digital tools available to aid linguistic research, notably absent are tools for managing semantic and, more broadly, linguistic analysis by keeping data organized, structured, and retrievable, and providing the flexibility to construct our own analytical frameworks. In this note, we consider the advantages offered by relational databases in this respect and relate the experience of a lexicographical project. We review relational databases'

main features and workings, describe the design of the one used in said project, and present the benefits yielded by using this tool.

Keywords:
relational databases, semantics, language analysis, lexicography, digital tools

1. Introducción

Las tecnologías digitales son hoy ubicuas en todas las etapas de la investigación científica, desde la recolección de los datos hasta la publicación de los resultados. La lingüística no es la excepción. Como señalan Teira y Polo (2021),

> la Lingüística [*sic*] se ha convertido en una disciplina interdisciplinar que necesita nuevos materiales, herramientas y métodos, porque la tecnología ha permitido acceder a una gran cantidad de información que necesita formas nuevas de manejarla. (p. 158)

Las autoras presentan un panorama bien nutrido de recursos y herramientas digitales útiles en una pluralidad de ramas de la lingüística. Incluyen corpus, bases de datos lexicales y de otros tipos, repositorios, diccionarios, lematizadores, analizadores estadísticos, aplicables en estudios de morfología, sintaxis, fonética y fonología, lexicología y semántica.

En este último campo, otras propuestas incluyen, p. ej., DispoGrafo (Echeverría *et al.*, 2008), una herramienta para visualizar relaciones lexicales en un corpus, en forma de grafos que representan redes semánticas; y la recopilación de herramientas para compilar diccionarios que presentan Rubio López *et al.* (2021). Teira y Polo (2021) refieren también dos analizadores semánticos que funcionan a partir de corpus textuales.

Sin embargo, estos analizadores semánticos trabajan sobre su propia ontología y su propio sistema de rasgos, ya predeterminados. Por su parte, las herramientas lexicográficas que recogen Rubio López *et al.* (2021) están diseñadas para facilitar el trabajo de edición de diccionarios, una vez que ya el análisis semántico prelexicográfico ha sido realizado. Se hace notar entonces la ausencia de herramientas para gestionar el análisis semántico —no solo con fines lexicográficos—, que nos permitan a los analistas mantener la información organizada y recuperable durante el proceso de investigación, en especial en proyectos que implican el análisis de corpus voluminosos, y que nos brinden la flexibilidad para construir nuestros propios aparatos analíticos en función de nuestro marco teórico.

El objetivo de esta nota es exponer las bondades que ofrece al análisis lingüístico, específicamente semántico, otro tipo de herramienta digital: las bases de datos relacionales (BDR); y referir como caso ilustrativo nuestra propia experiencia en un proyecto lexicográfico.

Ofrecemos, en la sección 2, un panorama general de las BDR, sus características y funcionamiento. La sección 3 refiere nuestra experiencia: el diseño de y trabajo con la base de datos y los beneficios que hemos obtenido con el uso de esta herramienta digital. En conclusión, consideramos las ventajas generales y limitaciones de su aplicación a los análisis lingüísticos.

2. Bases de datos relacionales

Una base de datos es una colección organizada de datos para un propósito específico (Hernandez, 2013, cap. 1, sec. «Types of Databases», párr. 1). Existen dos tipos: las bases de datos operacionales, utilizadas por muchas empresas y organizaciones, que almacenan datos dinámicos (que cambian constantemente) en tiempo real y de las cuales depende el funcionamiento de la institución (párr. 3); y las bases de datos analíticas, que almacenan datos estáticos (que no cambian o lo hacen muy poco) y permiten, p. ej., establecer tendencias y obtener otras informaciones estadísticas (párr. 4). Para un proyecto de investigación lingüística, estaríamos hablando de una base de datos analítica, pues en ella almacenaríamos y organizaríamos los datos recolectados durante el análisis, para luego obtener de ellos la información que nos interese.

La manera en que se almacenan y organizan los datos, es decir, la estructura de la base de datos, recibe el nombre de modelo de datos o intensión, y se opone a la extensión o estado de la base datos, que denota los datos que esta contiene efectivamente en un momento dado. El modelo de datos es así una estructura abstracta que define qué datos es posible almacenar en la base de datos y cómo (Ricardo, 2009, p. 70). Desde 1970 y entre los diversos modelos de datos existentes, el modelo relacional es aún el más ampliamente utilizado (Hernandez, 2013, cap. 1, sec. «The Relational Database Model», párr. 1; Ricardo, 2009, p. 72).

El modelo relacional[1] debe su nombre al término matemático «relación», utilizado en teoría de conjuntos, en el cual no nos detendremos aquí. En este modelo, los datos se almacenan en tablas cuyas filas se denominan tuplas o registros, y cuyas columnas reciben el nombre de atributos o campos. Por ejemplo, en una tabla de verbos, cada registro podría corresponder a un verbo y contener campos como «Verbo», «Valencia», «ArgumentoSujeto», «ArgumentoObjeto», etc. El orden físico en que aparezcan registros y

[1] Una introducción accesible puede encontrarse en IONOS (2020).

campos es irrelevante y cada registro contiene, en un campo determinado, un valor único que lo identifica como distinto de los otros (Hernandez, 2013, cap. 1, sec. «The Relational Database Model», párr. 3), a menudo un identificador numérico generado automáticamente para cada nuevo registro y almacenado en un campo llamado «Id». Este identificador recibe el nombre de clave primaria.

Las tablas de una BDR pueden estar relacionadas entre sí, hecho que se toma a menudo y erróneamente como origen del término «relacional». Por ejemplo, la anterior tabla de verbos puede estar relacionada con una tabla de roles semánticos. Esta tabla contendría campos como «Id» (clave primaria) y «Rol», y almacenaría registros como «Agente», «Paciente», «Locativo», etc. En la tabla de verbos, el campo «ArgumentoSujeto» conectaría con el campo «Rol» de la tabla de roles semánticos, que sería entonces la clave externa de la tabla de verbos. Esta simple relación sería ya suficiente para obtener, digamos, todos los verbos de la base de datos que toman un sujeto agente. Lo mismo es aplicable al campo «ArgumentoObjeto», lo cual crearía una segunda relación.

Las relaciones entre las tablas pueden ser de varios tipos. A un registro en la tabla A puede corresponder un único registro en la tabla B o varios; a varios registros en la tabla A puede corresponder un mismo y único registro en la tabla B o varios. En nuestro ejemplo, a varios registros en la tabla de verbos puede corresponder un mismo y único registro en la tabla de roles semánticos, pues son varios los verbos que toman un sujeto agente, pero cada verbo (en el ejemplo que usamos) tiene un único argumento sujeto con un único rol semántico.

En el ejemplo, las dos tablas se relacionan directamente. Sin embargo, también es posible crear relaciones indirectas, a través de tablas intermedias. Supongamos que tenemos una tercera tabla de clases ontológicas, con los campos «Id» y «Clase», que almacena valores como «Humano», «Inanimado», «Propiedad», etc. Podemos relacionarla con la tabla de roles semánticos, donde crearíamos un campo «Clase» conectado al campo homónimo de la nueva tabla de clases ontológicas. Esta relación nos diría qué clases ontológicas pueden asumir qué roles semánticos[2]. Puesto que ahora la tabla de clases ontológicas se relaciona indirectamente con la tabla de verbos, por medio de la tabla de roles semánticos, también podemos saber, p. ej., qué verbos toman un sujeto humano o un objeto inanimado. Aclaramos que esta es solo

[2] Para simplificar, estamos postulando que un rol semántico puede corresponder a una única clase ontológica y viceversa. No obstante, también es posible crear relaciones de varios a varios.

una de las maneras en que se podrían organizar estos datos. El modelo relacional es sumamente flexible y la decisión dependerá del analista.

No existe límite teórico para la cantidad de tablas y de relaciones que pueden conformar una base de datos. La estructura puede ser tan compleja como lo requiera el análisis. Siempre que el analista conozca las tablas y las relaciones entre ellas, «podrá acceder a los datos de maneras casi ilimitadas» (Hernandez, 2013, cap. 1, sec. «The Relational Database Model», párr. 5. Trad. del autor).

Ahora bien, ¿qué herramientas existen para crear y trabajar con BDR? En nuestra experiencia, los sistemas de gestión de bases de datos relacionales (SGBDR) más accesibles a un lingüista sin formación especializada en bases de datos son los SGBDR para escritorio, tales como los sistemas abiertos LibreOffice Base y Calligra KEXI, o el comercial Microsoft Access, aunque existen otros. La elección dependerá de la disponibilidad y las preferencias del usuario. Es importante señalar que los softwares de hojas de cálculo como LibreOffice Calc, Calligra Sheets o Microsoft Excel no son SGBDR, aunque la información se almacene también en tablas.[3]

Hemos hablado de almacenar los datos y de recuperarlos. Esta recuperación se realiza por medio de consultas a la base de datos, donde se especifican los criterios para los datos que se desea recuperar. En estas consultas (y también en la creación y modificación de las bases de datos), los SGBDR utilizan el *Structured Query Language* (SQL), un lenguaje de programación que, en 1986, fue adoptado por el American National Standards Institute (ANSI) y por la International Standards Organization (ISO) como lenguaje estándar para BDR (Ricardo, 2009, p. 210). Dicho esto, diferentes SGBDR pueden modificar ligeramente la sintaxis de SQL e incorporarle sus propias extensiones, así que es posible hablar de «dialectos» de SQL, aunque los fundamentos se mantienen siempre.

Para los no especialistas, un lenguaje de programación introduce un nuevo nivel de complejidad en el trabajo con bases de datos. Sin embargo, coincidimos con Teira y Polo (2021) cuando afirman que

> [l]os lenguajes de programación no deberían ser ajenos a nuestra profesión (ej. R o Python son algunos de ellos y su conocimiento permite aprovechar al máximo las posibilidades de análisis

[3] Puede verse una comparación punto por punto entre hojas de cálculo y bases de datos en KDE (2012).

> de grandes conjuntos de datos). Los lingüistas necesitamos dominar cada vez más estos programas para entender los resultados y poder replicar estudios. (p. 171)[4]

No es nuestro objetivo detenernos en la estructura y funcionamiento de SQL.[5] Ofrecemos, a modo ilustrativo, un ejemplo de una consulta a la base de datos de verbos que hemos venido utilizando como ejemplo:

```
SELECT Id, Verbo FROM Verbos
WHERE ArgumentoSujeto = 'Agente'
ORDER BY Verbo;
```

Esta consulta le dice a la base de datos que recupere (SELECT) los campos «Id» y «Verbo» de (FROM) la tabla «Verbos», para los registros que cumplan con una condición (WHERE): el campo «ArgumentoSujeto» debe tener el valor (=) «Agente»; y que ordene los resultados (ORDER BY) alfabéticamente por el campo «Verbo». La consulta cierra con un punto y coma (;) y, en respuesta, devuelve una tabla con los datos solicitados.

Estos son los tres componentes de una consulta básica en SQL: SELECT…FROM, WHERE y ORDER BY (Hernandez, 2013, cap. 1, sec. «Retrieving Data», párr. 2). Como se observa, no es un lenguaje de programación sumamente enrevesado (aunque tiene sus complejidades) y respeta una lógica que es fácil de seguir.

Es cierto que, si utilizamos un SGBDR como los mencionados antes, que ofrecen una interfaz gráfica para construir consultas, nunca tendremos que escribir código SQL. Sin embargo, como señala Hernandez (2013), conviene entender los rudimentos, pues nos permitirá solucionar cualquier problema que surja con las consultas creadas a través de la interfaz gráfica (cap. 1, sec. «Retrieving Data», párr. 4).

En resumen, entonces, el modelo relacional ofrece una estructura básica simple, intuitiva y fácil de comprender: datos almacenados en tablas con registros y campos, que se moldean de acuerdo con las necesidades. Las tablas pueden relacionarse entre sí siguiendo la lógica propia de los datos y del analista. Es fácil operar con los datos y el lenguaje de programación

---

[4] Levshina (2015) presenta una guía muy completa para el uso de R en lingüística. Para Python, puede consultarse Hammond (2020).

[5] Un tutorial básico y accesible está disponible en «Tutorial de SQL» (s. f.).

SQL brinda una gran potencia con un conjunto pequeño de comandos. No en balde el relacional sigue siendo, más de medio siglo después, el modelo de datos más popular.

3. La experiencia de DIVERCE

La experiencia que compartimos aquí consiste en el uso de una BDR para gestionar el análisis semántico constitutivo de un proyecto lexicográfico en curso: el Diccionario Semántico de Verbos Causativos del Español (DIVERCE). Se trata de un proyecto complejo, por el volumen tanto del corpus como de los datos que se obtienen del análisis, lo cual motivó la búsqueda de una solución logística para la gestión del análisis. Otros modos de organización quizás más tradicionales —p. ej., matrices semánticas almacenadas en documentos de texto o en hojas de cálculo— resultaron ser, desde un inicio, insuficientes.

DIVERCE es un proyecto de diccionario electrónico[6] semántico, es decir que «se propone dar una presentación sistemática de los significados» a partir de «una teoría semántica que [brinde] un aparato analítico adecuado» (Rasmussen, 2014, p. 39). Dicha teoría determinará «bajo cuál perspectiva se debe indagar el problema» y «cuáles informaciones se actualizarán en la definición» (Piedra Matamoros, 2021, pp. 177-178). DIVERCE toma como base el modelo semántico de la causalidad propuesto por Wong García (2020a y 2020b), del cual presentaremos los elementos mínimos indispensables según sean necesarios.

Es también un proyecto de diccionario onomasiológico, cuya direccionalidad va de los significados más o menos generales a los significantes que los expresan. Esto exige una organización estructurada de los datos que permita al usuario final construir de manera dinámica su recorrido lexicográfico y obtener los resultados que busca.

Por último, un diccionario de este tipo no es solamente un producto lexicográfico; es también un producto teórico, una descripción sistemática de un campo semántico. En esto se asemeja a cualquier otro análisis lingüístico, pues los datos que almacena deben estar organizados de tal manera que permitan operar con ellos, establecer generalizaciones, formular y verificar hipótesis, y producir conocimiento.

---

[6] «Un diccionario electrónico es un diccionario que no existe previamente en versión impresa, cuyo diseño implica la creación de una base de datos *ad hoc* en función de los contenidos que se desea incluir, los cuales se vuelcan de modo manual o semi-manual, y son consultados en línea mediante una interfaz.» (Barrios Rodríguez, 2020, p. 36)

3.1. Diseño de y trabajo con la base de datos

Apuntábamos en la introducción que las herramientas lexicográficas existentes (Rubio López *et al.*, 2021) están diseñadas para facilitar el trabajo de edición del diccionario, no para gestionar la labor prelexicográfica; además, están pensadas para diccionarios semasiológicos tradicionales. Así que, para un proyecto de diccionario semántico onomasiológico como DIVERCE, que exige un intenso trabajo de análisis prelexicográfico que es indispensable gestionar, debimos recurrir a otras herramientas. Puesto que la elaboración de un diccionario electrónico implica el diseño de una base de datos (Barrios Rodríguez, 2020, p. 36), no vimos la necesidad de esperar a culminar el trabajo prelexicográfico para diseñarla, cuando la misma base de datos podía utilizarse en el análisis. Como es el modelo de datos más ampliamente utilizado, decidimos que sería una BDR.[7]

Hernandez (2013) recomienda diseñar la estructura de las bases de datos sin pensar en ningún SGBDR en particular, para que el conocimiento que se tiene de las funcionalidades de un software específico no condicione el diseño (cap. 14, sec. «Database Design Based on the Database Software», párr. 9). Esto es aplicable sobre todo a los diseñadores profesionales de bases de datos. Sin embargo, un lingüista que, en la práctica, necesita una herramienta para gestionar su trabajo, inevitablemente basará su diseño en el SBGDR más fácilmente disponible. Fue nuestro caso. Trabajamos con Microsoft Access, aunque cualquier SBGDR de los mencionados antes permitirá obtener los mismos resultados (salvando quizás algunas funcionalidades muy específicas), pues todos comparten el modelo relacional.

Para el diseño de nuestra base de datos, partimos de la distinción que encontramos en Ricardo (2009, pp. 51-53) entre los cuatro niveles de abstracción al trabajar con datos:

1. El mundo real y, dentro de él, la parte que se representará en la base de datos, llamada minimundo o universo de discurso.
2. El modelo conceptual, formado por las entidades, clases de entidades, atributos y relaciones identificados en el mundo real.

[7] Durante el proceso, hemos experimentado con otros tipos de bases de datos no relacionales, sobre todo con bases de datos por grafos (ing. *graph databases*). Sin embargo, los mejores resultados los hemos obtenido con el modelo relacional.

3. El modelo lógico, la estructura de la base de datos, que incluye los tipos de datos que se van a almacenar (tablas y campos).
4. Las instancias de datos, los datos reales almacenados en la base de datos.

No es difícil ver que estos cuatro niveles de abstracción se alinean con los componentes de un análisis lingüístico. El universo de discurso (1) corresponde al objeto de estudio; el modelo conceptual (2), al aparato teórico; el modelo lógico (3), al instrumento analítico; y las instancias de datos (4), a los resultados del análisis. En nuestro caso, el universo de discurso lo constituía la estructura semántica de los verbos causativos del español. Contábamos también con un primer modelo conceptual, tomado de la teoría de base (Wong García, 2020a y 2020b).

Decimos «un primer modelo conceptual» porque este proceso de abstracción no es lineal. Pasar del nivel 1 al 2 conlleva necesariamente un análisis, aunque sea preliminar, y el estudio subsiguiente del objeto en el nivel 4 va a producir modificaciones en el modelo conceptual que se traducirán en cambios en el nivel 3. Por esta razón, realizamos primero un análisis exploratorio con una muestra pequeña de 50 verbos causativos frecuentes, con dos objetivos: verificar la aplicabilidad de la teoría de base e identificar los cambios y extensiones que habría que hacerle para transformarla en el modelo conceptual de nuestra base de datos. En efecto, tuvimos que extenderla para hacerle ganar en precisión y adaptarla a la semántica lexical, pues había sido formulada originalmente para el estudio de unidades discursivas.

En síntesis, combinando la teoría de base y las demás informaciones lexicográficas que debe brindar el diccionario, el modelo conceptual quedó estructurado como sigue:

- Los lemas verbales tienen un significante y pueden tener una o más acepciones o variantes léxico-semánticas (VLS) causativas. Estas VLS tienen un orden dentro del lema, un modo de significación (directo, metafórico, extensión, restricción, etc.), marcas diasistémicas (peyorativo, despectivo, dominios de uso, etc.), una definición, precisiones gramaticales ocasionales y ejemplos. En cada una de estas VLS, se construye lo que en Wong García (2020a y 2020b) se denomina escenarios causales (EC) y que hemos llamado aquí configuraciones.
- Cada una de estas configuraciones pertenece a un género de causalidad (hacer-devenir, hacer-hacer, hacer-experimentar y otros), a un tipo de EC (causación, impedimento, permisión y otros), en ocasiones a un subtipo de EC, a un campo

nocional (espacial, temporal, afectivo-emocional y otros) y, en ocasiones, a un subcampo nocional.

- Cada una de estas configuraciones tiene una estructura argumental. Sus argumentos pertenecen a clases ontológicas (humano, inanimado, objeto físico, etc.), asumen roles semánticos (agente, paciente, tema, etc.) y algunos se realizan como constituyentes sintácticos (sujeto, primer complemento, segundo complemento, etc.), introducidos o no por un nexo gramatical. Otros están incorporados al verbo y otros tienen realización sintáctica opcional.
- Estos argumentos pueden recibir modificadores (p. ej., atributos) y, sobre todo en VLS complejas con más de una configuración (piénsese en el verbo *intimidar*, que construye simultáneamente un EC de «obligar» y uno de «causar miedo»), pueden ser correferenciales entre ellos.
- Las VLS también deben remitir, dentro del diccionario, a otras de significado cercano.

Con este modelo conceptual, ya disponemos de una lista de las clases de entidades, de sus atributos y de las relaciones entre ellas:

> Clases de entidades — lemas, VLS, configuraciones, géneros de causalidad, (sub)tipos de EC, (sub)campos nocionales, argumentos, clases ontológicas, roles semánticos, modificadores, constituyentes sintácticos.
>
> Atributos — significante del lema, orden de las VLS, modo de significación, marcas diasistémicas, definición, notas gramaticales, ejemplos, nexo gramatical, carácter opcional e incorporado.
>
> Relaciones — ser VLS de un lema, ser configuración de una VLS, pertenencia (a un género de causalidad, a un tipo de EC, a una clase ontológica…), ser argumento de una configuración, asumir un rol, realización sintáctica, modificación, correferencia, remisión.

En principio, entonces, las clases de entidades se convierten en tablas; los atributos, en campos; y las relaciones, en relaciones entre tablas. En la práctica, se pueden tomar otras decisiones. Por ejemplo, decidimos crear también una tabla para los modos de significación y una para las marcas diasistémicas, de manera que, al rellenar esos campos en la tabla de VLS, solo hubiera que escoger una opción de una lista, en lugar de escribirlos manualmente.

Decidimos también almacenar las remisiones en una tabla aparte, utilizando como referencias los identificadores numéricos únicos de las VLS. Esto tiene la ventaja de que, si se utiliza un SGBDR que no admite más de un valor por campo, se pueden almacenar remisiones de una VLS a varias, creando un registro para cada remisión individual.

Todas estas tablas están relacionadas entre sí, utilizando para ello las claves primarias y externas, tal como explicamos antes. Así, por ejemplo, la tabla de configuraciones contiene los campos «Id», «VLS» «Género», «EC», «SubEC», «CN» y «SubCN». El primero es el identificador numérico único para cada registro; el segundo conecta con el campo «Id» de la tabla de VLS (almacena a qué VLS pertenece una configuración dada); y los últimos cinco conectan con el campo «Id» de las tablas de género de causalidad, tipos de EC, subtipos de EC, campos nocionales y subcampos nocionales, respectivamente. Asimismo, la tabla de argumentos contiene, entre otros, un campo «Config», que conecta con el campo «Id» de la tabla de configuraciones (almacena en qué configuración participa un argumento dado). Por cuestiones de espacio, no reproducimos aquí la integralidad de las tablas, sus campos y sus relaciones.

A la hora de ingresar datos en la base de datos, el analista tiene dos opciones. Puede trabajar directamente con sus tablas «activas» (en las que se introduce la información, opuestas a las «pasivas», que almacenan datos más fijos) o puede crear formularios personalizados. La segunda opción es más cómoda, pero, para bases de datos de cierta complejidad como la nuestra, exige habilidades para poder incluir en un mismo formulario todas las tablas relacionadas. La ventaja de Microsoft Access es que permite, al trabajar directamente con las tablas, subordinarlas unas a otras, de manera que, desde la tabla de lemas (en la cima de jerarquía), se puede entrar información en todas las demás; para cada registro, el sistema anida dentro de él los registros correspondientes de la tabla inmediatamente inferior: la de VLS; luego, la de configuraciones; luego, la de argumentos; y, por último, la de modificadores de argumentos. Esta jerarquía se genera automáticamente a partir de las relaciones que se han definido entre las tablas, pero el sistema permite también especificar, para cada tabla, cuál será su tabla secundaria o subordinada. La correspondencia entre los registros de una tabla y otra se garantiza por medio de las claves primarias y externas, es decir, los campos en los que se basa la relación entre las tablas.

3.2. Beneficios

Los beneficios de gestionar el análisis semántico con una BDR y de estructurar esta como hemos descrito, se hacen ver en la organización y estructuración de los datos, en la recuperación de información por medio de consultas, en la presentación de la información y en las futuras aplicaciones de la base de datos.

En términos de la organización de los datos, damos por sentado lo conveniente de almacenarlos de manera estructurada y no plana como, p. ej., en una matriz semántica, donde datos como la estructura argumental serían difícilmente representables. Por otra parte, utilizar tablas para almacenar datos constantes, como los (sub)tipos de EC, los (sub)campos nocionales o las marcas diasistémicas, y relacionar estas con las tablas activas por medio de claves primarias y externas, no solo economiza la entrada de datos, sino que además garantiza que cualquier modificación que sea necesario hacer se refleje automáticamente en toda la base de datos. Por ejemplo, si decidiéramos cambiar el nombre que utilizamos para un tipo de EC, no tendríamos que cambiarlo en todos los registros uno a uno; bastaría con modificar el registro correspondiente en la tabla de tipos de EC y se actualizarían todos.

Durante el análisis, pueden surgir nuevos datos que haya que incorporar al modelo conceptual. En un proyecto lexicográfico como el nuestro, pueden sumarse también informaciones lexicográficas que se desee brindar (como fue el caso del modo de significación, no incluido originalmente). En estos casos, una BDR permite añadir tablas nuevas y campos nuevos a las ya existentes, y crear nuevas relaciones entre ellos, de manera simple y sin que se afecte la integridad de los datos ya almacenados.

En lo que respecta a la recuperación de información, quizás los mayores beneficios sean los aportados por las consultas estadísticas. Cuando se trabaja con corpus voluminosos que producen numerosas instancias de datos, es fácil pasar por alto generalizaciones cuantitativas; una BDR, aunque no brinda esta información automáticamente (más allá de la cantidad de registros que contiene cada tabla), facilita su obtención.

La información estadística más simple que podemos obtener es la que se refiere a una única tabla; p. ej., todos los registros que contienen determinado valor en un campo, digamos, todas las configuraciones de género «hacer-devenir». Para esto, en realidad no hace falta crear una consulta. Los SGBDR permiten aplicarles filtros a las tablas para mostrar solamente ciertos registros. No obstante, por medio de una consulta, se puede obtener el número total,

en lugar de los registros individuales, utilizando la función Count de SQL. Sin embargo, las relaciones entre las tablas de la BDR permiten crear consultas que arrojen información estadística mucho más relevante y difícil o trabajosa de obtener por otras vías.

Así, en DIVERCE, estamos llevando cuenta de los índices de polisemia (cuántas VLS por lema, aprovechando la relación entre las tablas de lemas y de VLS); de la distribución de las configuraciones por género de causalidad, por (sub)tipo de EC y por (sub)campo nocional (aprovechando las relaciones de la tabla de configuraciones con las tablas correspondientes); de la complejidad estructural de las configuraciones (cuántos argumentos por configuración, aprovechando la relación entre las tablas de configuraciones y de argumentos); de las diez clases ontológicas más frecuentes en las estructuras argumentales (contando los totales de cada una en la tabla de argumentos, ordenando de mayor a menor y limitando la consulta a los diez primeros registros); y de los diez roles semánticos más frecuentes (utilizando el mismo método).

Podemos también, según necesidades puntuales, obtener todas las VLS en las que se realice determinado rol semántico (sirviéndonos de la relación indirecta «tabla de VLS > tabla de configuraciones > tabla de argumentos > tabla de roles semánticos») o en las que ocurra determinada clase ontológica en determinado constituyente sintáctico (utilizando las relaciones indirectas «tabla de VLS > tabla de configuraciones > tabla de argumentos > tabla de clases ontológicas» y «[…] tabla de argumentos > tabla de constituyentes»).

Decíamos que los principales SGBDR para escritorio ofrecen una interfaz gráfica para la creación de consultas, por lo cual no es imprescindible conocer SQL. Sin embargo, se debe tener presente que, mientras mayor sea la complejidad de la consulta, mayor será también la probabilidad de que el analista necesite introducir directamente código SQL o modificar el que genera el sistema. Si dominamos los rudimentos de SQL, podemos diseñar consultas más específicas y complejas, e identificar regularidades que, de otra forma, no habríamos notado.

Toda esta información estadística puede tomarse en forma de los datos brutos que produce la consulta o utilizar esa consulta como base para otros modos de presentación de la información. Los principales SGBDR permiten también crear informes, en los que se presentan los datos en un formato más accesible que una tabla, definido y personalizado por el usuario. Estos informes pueden incluir gráficos (de barras, de dispersión, circulares, etc.), que visualizan la información estadística. Teira y Polo (2021, pp. 170-171) citan algunas

herramientas para elaborar gráficos, pero no podemos negar lo conveniente de poder generarlos en el mismo sistema donde tenemos almacenados los datos. Estos gráficos pueden insertarse también en los formularios, según prefiera el analista. En caso de que las funcionalidades de gráficos no fueran suficientes, los SGBDR brindan también la posibilidad de exportar los datos en formatos normalizados, para poder utilizarlos en otros sistemas. Por lo demás, un informe puede contener información de una tabla o consulta, o de varias, lo cual hace posible combinar datos para presentar la información más relevante o necesaria en un momento dado.

Finalmente, gestionar el análisis semántico en una BDR nos va a permitir, al concluir el proyecto, utilizar esta misma base de datos para desplegar el diccionario en línea. Puesto que ha ido creciendo y estructurándose a la par que el corpus analizado, no habrá necesidad de utilizar tiempo y recursos diseñando una base de datos para la aplicación en línea. Para proyectos no lexicográficos, esto significa que los datos recopilados, que a menudo se dejan de lado una vez culminada la investigación, pueden ponerse a disposición de otros, en un formato organizado y estructurado, para que los consulten y continúen trabajando con ellos.

4. Conclusiones

Hemos reseñado las características generales y el funcionamiento de las BDR, y referido nuestra propia experiencia con su aplicación a la gestión del análisis semántico constitutivo de un proyecto lexicográfico: cómo hemos diseñado la base de datos y los beneficios que esta herramienta nos ha reportado.

Entre las principales ventajas, se cuenta el hecho de que, al contrario de los analizadores semánticos disponibles en Internet, utilizar una BDR permite elaborar nuestro propio aparato analítico, según la teoría de base y las necesidades de la investigación. Permite, además, almacenar los datos resultantes de manera organizada, estructurada y fácilmente recuperable.

Estas herramientas ofrecen también la ventaja de poder obtener información estadística donde mismo se almacenan los datos, en lugar de recurrir a una herramienta externa. La complejidad y precisión de esta información estadística dependerá de la complejidad de las consultas en las que se basan.

Aunque nos hemos centrado en el análisis semántico, las BDR pueden ser de utilidad para gestionar cualquier tipo de análisis lingüístico que implique corpus voluminosos: desde análisis fonológicos o morfológicos, hasta análisis de discurso.

Quizás la principal limitación sea la curva de aprendizaje, tanto en lo que respecta al manejo de un SGBDR desconocido como al uso de SQL para diseñar consultas. No obstante, nuestra experiencia nos lleva a creer que los beneficios de contar con una herramienta para gestionar el trabajo analítico compensan con creces las dificultades técnicas.

BIBLIOGRAFÍA

Barrios Rodríguez, M. A. (2020). «¿Aún queda alguien para quien no exista un diccionario? Diretes, un diccionario electrónico apto para máquinas», en M. C. Cazorla Vivas, M. A. García Aranda, y M. P. Nuño Álvarez (Eds.), *Lo que hablan las palabras. Estudios de lexicografía y gramática en honor de Manuel Alvar Ezquerra*, pp. 33-46. Lugo: Axac.

Echeverría, M. S., Vargas, R., Urzúa, P., y Ferreira, R. (2008). «DispoGrafo: Una nueva herramienta computacional para el análisis de relaciones semánticas en el léxico disponible». *Revista de Lingüística Teórica y Aplicada*, *46*(1), 81-91. Disponible en https://www.scielo.cl/pdf/rla/v46n1/art05.pdf

Hammond, M. (2020). *Python for Linguists*. Cambridge: Cambridge University Press.

Hernandez, M. J. (2013). *Database Design for Mere Mortals. A Hands-On Guide to Relational Database Design* (3ra edición) [Libro electrónico]. Nueva York: Pearson Education. Disponible en https://www.ebooks.com/en-bm/book/1126748/database-design-for-mere-mortals/michael-j-hernandez/

IONOS. (2020, enero 8). «Bases de datos relacionales: El modelo de datos en detalle». Sitio web: https://www.ionos.es/digitalguide/hosting/cuestiones-tecnicas/bases-de-datos-relacionales/ (17/08/2021)

KDE. (2012, septiembre 14). «Las bases de datos y las hojas de cálculo» (R. González y J. M. García Molina, Trads.). Sitio web: https://docs.kde.org/trunk5/es/kexi/kexi/database-and-spreadsheet.html (17/08/2021)

Levshina, N. (2015). *How to do Linguistics with R. Data exploration and statistical analysis.* Ámsterdam, Filadelfia: John Benjamins.

Piedra Matamoros, E. (2021). «Generación de definiciones para un diccionario escolar de la sexualidad». *Revista Pensamiento Actual*, *21*(36), 172-180. https://doi.org/10.15517/PA.V21I36.47022V

Rasmussen, S. (2014). «Los aspectos metalexicográficos de un diccionario semántico. Reflexiones en torno a un diccionario semántico de los verbos españoles», en G. Wotjak (Ed.), *Estudios de lexicología y metalexicografía del español actual*, pp. 38-62. Berlín, Boston: Max Niemeyer Verlag.

Ricardo, C. M. (2009). *Bases de datos* (1ra edición en español; V. Campos Olguín y J. Enríquez Benito, Trads.). México, D. F.: McGraw Hill.

Rubio López, R. Y., Estiven Bonilla, J., y Bernal Chávez, J. A. (2021). «*Dictionary Writing Systems* y otras herramientas informáticas para la elaboración, administración y publicación de diccionarios». *Lingüística y Literatura*, *80*, 340-360. https://doi.org/10.17533/udea.lyl.n80a20

Teira, C., y Polo, N. (2021). «Digitalización y recursos para la investigación en lingüística». *Revista Española de Lingüística*, *51*(1), 157-176. https://doi.org/10.31810/rsel.51.1.9

Tutorial de SQL. (s. f.). Sitio web: https://desarrolloweb.com/manuales/tutorial-sql.html (16/08/2021)

Wong García, E. (2020a). «Aspectos de un modelo semántico de la causalidad». *Signos Lingüísticos*, *XVI*(31), 8-43. Disponible en https://signoslinguisticos.izt.uam.mx/index.php/SL/article/view/252

Wong García, E. (2020b). *Causalidad y modalidad: Un modelo semántico-discursivo de la causalidad para el análisis del discurso modalizado* (Tesis de doctorado). Universidad de La Habana. http://dx.doi.org/10.13140/RG.2.2.21048.78088