# La densidad lexicométrica: del *big data* a la evidencia lingüística

**Jorge Lázaro**
Universidad Autónoma de Baja California
*lazaroj@uabc.edu.mx*

**Llicenciat en llengua i literatures** hispàniques per la Facultat de Filosofia i Lletres de la Universitat Nacional Autònoma de Mèxic (UNAM) i doctor en comunicació lingüística i mediació multilingüe per la Universitat Pompeu Fabra. Membre del Grup d'Enginyeria Lingüística, on va ser investigador postdoctoral (2015-2016). Ha estat professor a la Facultat de Filosofia i Lletres i a la Facultat d'Enginyeria de la UNAM. Actualment és professor investigador titular a la Facultat d'Idiomes de la Universitat Autònoma de Baixa Califòrnia. És membre del Sistema Nacional d'Investigadors del Consell Nacional de Ciència i Tecnologia de Mèxic (CONACyT). Les seves línies de recerca són la lingüística computacional, la terminologia, la lexicologia, la lexicografia computacional i la semàntica.

**Resum**

***La densitat lexicomètrica: del* big data *a l'evidència lingüística***

*El present article aborda el* big data, *vist des de la perspectiva de la lingüística de corpus, en el que s'anomena* corpus massius. *A través de la revisió d'alguns treballs i d'un estudi de cas, es fa evident el canvi de paradigma en la lingüística aplicada actual. Es presenta una proposta d'extracció automàtica d'exemples com a prova d'aquesta afirmació.*
PARAULES CLAU: big data; *densitat lexicomètrica; saturació semántica; exemplificació*

**Resumen**

*El presente artículo aborda el* big data, *visto desde la perspectiva de la lingüística de corpus, en lo que se denomina* corpus masivos. *A través de la revisión de algunos trabajos y de un estudio de caso, se hace evidente el cambio de paradigma en la lingüística aplicada actual. Se presenta una propuesta de extracción automática de ejemplos como prueba de dicha afirmación.*
PALABRAS CLAVE: big data; *densidad lexicométrica; saturación semántica; ejemplificación*

**Abstract**

***Lexicometric density: from Big Data to linguistic evidence***

*This paper approaches Big Data from the perspective of corpus linguistics in what is called* massive corpora. *By the review of some research studies, and a case study, the paradigm shift in current applied linguistics is made evident. We present a proposal for automatic extraction of examples as proof of this affirmation.*
KEYWORDS: *Big Data, lexicometric density, semantic saturation, exemplification*

TERMINÀLIA 19 (2019): 17-27 · DOI: 10.2436/20.2503.01.130
Data de recepció: 15/3/17. Data d'acceptació: 1/10/2018
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · *http://terminalia.iec.cat*

## 1. Introducción

Los datos masivos (o *big data*) son desde hace algunos años una realidad indiscutible. Es cada vez mayor la tendencia a utilizar una cantidad inconmensurable de todo tipo de información para poder satisfacer necesidades tan distintas como el *marketing* o la salud, la física o la minería de opinión, la geografía o la lingüística, entre muchas otras áreas del conocimiento humano (Mayer-Schönberger y Cukier, 2013).

Sin embargo, debido a que el acceso a estos datos era hasta hace poco tiempo libre o poco regulado, también puede darse el caso de una mala utilización de la información almacenada para llevar a cabo ilícitos que desembocan algunas veces en delitos informáticos. Sin ir más lejos, el robo de información personal para acrecentar una cartera de clientes es frecuente e incluso se ha usado para incurrir en suplantación de identidad o *phishing*.[1]

En cada caso que analicemos en el que se usen datos masivos podremos encontrar por lo menos tres características:

1) Los datos masivos dan precisión a los estudios de todo tipo. Cuantos más datos disponibles, menor margen de error.
2) Los datos masivos permiten predecir un comportamiento. Una vez que puede cuantificarse un fenómeno, es mucho más fácil reproducirlo para después interpretar los resultados, agregando variantes.
3) Los datos masivos permiten determinar patrones. Hay fenómenos que a pequeña escala no presentan continuidad o recursividad, aun cuando la intuición de quien los estudia le dicta que dichos patrones están ahí porque en los sistemas macro, como el lenguaje, puede verlos repetidos.[2]

En la lingüística, los datos masivos se han asociado en los últimos años a áreas como el aprendizaje automático (*machine learning*) para entrenar programas derivados de la ingeniería lingüística que en conjunción con la lingüística de corpus han aportado repositorios bastante amplios. Puede hablarse de varios casos en los que los análisis lingüísticos, los corpus y las nuevas tecnologías han convergido para apoyar áreas emergentes o consolidar otras. Por ejemplo, el perfilado de autor (López-Escobedo, Solórzano-Soto y Sierra, 2016) o la clasificación y el resumen automático de textos (Torres-Moreno, 2014). Por otro lado, el material utilizado para realizar pruebas o directamente ciertos análisis nos adentra en una amplia discusión, debido a que, si bien hay una extensa tradición en cuanto al tratamiento de los bancos y repositorios de datos lingüísticos, llamados en general *corpus*, el uso que se hace de ellos y el procesamiento específico corresponden a las finalidades de cada investigación. Hay corpus de millones y millones de palabras que se usan para muchas tareas sencillas, como los consultadísimos CREA[3] (Corpus de Referencia del Español Actual) (más de 200 millones) y CORDE[4] (Corpus Diacrónico del Español) (250 millones), y hay corpus pequeños, de apenas unos cuantos millones de palabras, que se han explotado para varias investigaciones de gran envergadura, como en el caso del Corpus del Español de Mark Davies. Lo que queda claro en el debate sobre el tamaño de un corpus es que el *big data* ha permeado en lo más profundo de los análisis sobre el conocimiento humano y el estudio del lenguaje no es la excepción.

Si lo miramos al detalle, un corpus lingüístico podía considerarse grande si contenía 10 millones de palabras hasta hace apenas unos años (Torruella y Llisterri, 1999); sin embargo, el avance de la tecnología, y sobre todo de los medios de almacenamiento de datos en equipos de cómputo, ha permitido que en poco menos de dos décadas hayamos pasado de decir que un corpus es mínimo si contiene 1 millón de palabras (o pequeño si contiene 10 millones) a considerar que otro de 100 o 200 millones representa un tamaño estándar. Aunado a esto, cada día se han presentado mayores retos en los estudios lingüísticos, que han visto en los corpus con gran cantidad de datos una fuente invaluable para sostener preceptos teóricos o proponer nuevos paradigmas. Siguiendo esta tendencia, podríamos afirmar que en apenas unas décadas, desde que se ha generalizado su estudio y manejo, la lingüística de corpus nos ha enseñado que hay datos ocultos más allá de lo que el ojo humano puede ver a simple vista. Los casos de estudio, o muestras, aumentados en diez o veinte veces, pueden aportar pruebas fehacientes de hechos hasta el momento no comprobados (como veremos en el estudio de caso de este artículo) o descripciones lingüísticas más fiables (como mostraremos en la amplia lista de investigaciones).

## 2. Los corpus masivos

Muchos autores han hecho notar que el tamaño de un corpus siempre es un tema de discusión debido al carácter subjetivo de las afirmaciones que pueden hacerse con respecto al número de palabras o *tokens* que contiene (Torruella y Llisterri, 1999; Sardinha, 2000; Sierra, 2008; McEnery y Hardie, 2012), pero también es cierto que hay una gran diferencia entre los aproximadamente 10 millones que podemos procesar en casi cualquier equipo de cómputo doméstico y, por ejemplo, los casi 400 millones en un recurso como la *Wikipedia* en español.[5] Podemos encontrar repositorios que cuentan con más de 10.000 millones de palabras provenientes de todas las variantes de una lengua, como el corpus esTenTen —utilizado en Sketch Engine (Kilgarriff y Renau, 2013)— o la versión actual del corpus diseñado por Mark Davies —la versión web—, que contabiliza 2.000 millones de palabras. Cualesquiera que sean las razones, el procesamiento de grandes cantidades de datos es lo que nos lleva a integrar algu-

nos materiales bajo la denominación de *datos masivos*. Una de las características principales de este tipo de corpus es que no pueden ser visualizados en su totalidad por casi ningún procesador de texto. Es decir, este tipo de materiales de investigación deben visualizarse por fragmentos. Hablamos de *big data corpus* o de *corpus masivos* cuando su tratamiento se complica y requiere herramientas o recursos computacionales más avanzados y no pueden analizarse a simple vista.

Lo anterior ha llevado a los expertos lingüistas y computólogos a diseñar algoritmos para extraer o recuperar información relevante cada vez más precisos; la metodología de prueba-error entonces se hace necesaria hasta dar con el algoritmo ideal para una tarea lingüística. El filólogo ya no trabaja con ejemplos o casos específicos y comprueba hipótesis solo a través de estos, sino que ahora una hipótesis tiene que comprobarse con cientos, miles o millones de casos que pueden en cualquier momento mostrar una excepción a las generalizaciones que antes podían pasar de largo. Es decir, se busca el patrón de una hipótesis, no se generaliza a partir de casos específicos.

## 3. La evidencia lingüística y el big data en los estudios actuales

Un patrón en lingüística es un asunto complejo conceptualmente y es difícil de describir. Siempre puede haber excepciones a las posibles generalizaciones que se hacen acerca de una estructura (Carnie, 2013). Un patrón es una noción que, por un lado, puede parecer intuitiva y sistematizable, pero también resulta un elemento de trabajo casi interminable si tenemos en cuenta que los mismos actos de habla de cualquier persona representan siempre una iteración distinta. Aun así, podemos encontrar muchos patrones estables sobre todo en el nivel sintáctico o morfológico de la lengua. Ejemplos de patrones en lingüística son el número y el orden de los fonemas en una palabra, las declinaciones verbales en unas lenguas o las conjugaciones en otras, el orden de los elementos en una oración (SVO, sujeto-verbo-objeto; SOV, sujeto-objeto-verbo), la concordancia, etc. Sin embargo, hay otros patrones que se alejan un poco más de la sintaxis, se encuentran en el orden de lo sintáctico-semántico y resultan más difíciles de describir: la *consecutio temporum*, la ergatividad o el proceso de palatalización de una yod —hablando de fenómenos fonéticos e históricos—, etc. Dentro de estos patrones complejos encontramos también la estructura de las oraciones subordinadas referentes a ciertas funciones o la relación y estructura de las oraciones introducidas por marcadores discursivos (Cunha *et al.*, 2012).

Mientras más común es un elemento lingüístico, y, por tanto, más ejemplos de este existen en la lengua, menos datos son necesarios para poder transmitir o describir su mecanismo de funcionamiento. Pues bien, es justo esa cantidad de datos que se necesitan para documentar un fenómeno lo que hace compleja una tarea de caracterización (determinar un patrón para después reproducirlo).

El objetivo central de esta discusión es mostrar que ciertos fenómenos en la lengua tienen su base en patrones que no pueden ser aprehendidos fácilmente. Estos fenómenos o procedimientos necesitan muchos datos para ser caracterizados y se requiere un mayor número de ejemplos, es decir, su objeto de estudio demanda una gran cantidad de iteraciones para conocer todas sus variantes.

En la tabla 1 podemos observar los tamaños de algunos corpus europeos y orientales que están utilizándose en la actualidad para analizar fenómenos que tienen que ver sobre todo con el léxico, la lexicografía, la morfología y la sintaxis en diferentes lenguas.[6]

## 4. Un estudio de caso: la *densidad lexicométrica* como evidencia lingüística basada en el *big data*

Una investigación hecha previamente (Lázaro, 2015) ha demostrado que elementos lingüísticos como los ejemplos son estructuras que tienen poco que ver con un patrón sintáctico, pero que hay un trasfondo semántico que permite identificar algunas estructuras que se relacionan con definiciones especializadas. Esto ha sido posible a partir de ciertas técnicas de recuperación de información aplicadas a tres corpus masivos. Para llegar a este punto, tuvieron que hacerse, primero, una serie de presuposiciones lingüísticas basadas en la experiencia terminológica y apoyadas en la teoría comunicativa de la terminología (TCT) (Cabré, 1999) y en la noción de contextos definitorios (Sierra *et al.*, 2008). Después, tuvo que diseñarse una herramienta adecuada para demostrar dichas presuposiciones a partir de la puesta en escena de un elemento medible y comprobable. Sobre el plano teórico, puede consultarse el texto correspondiente a la propuesta denominada *saturación semántica* (Lázaro, 2016). La noción que ha intentado comprobar directamente dichas suposiciones a partir de técnicas propias de la lingüística computacional es la que se presenta a continuación: la *densidad lexicométrica*.

### 4.1 El ejemplo y un pequeño esbozo sobre su tratamiento actual

Desde el punto de vista teórico, el *ejemplo* ha sido visto por lexicógrafos y lingüistas de acuerdo con las necesidades del propio diccionario o como una exigencia de diseño. En el caso del castellano, hacia 1726, el *Diccionario de autoridades* era un modelo a seguir y su pilar más fuerte era justamente la inclusión de «autoridades» o citas, como las llamamos hoy. Desde 1780, cuando se convierte en el estandarte de la institución, la supresión de la mayoría de estas citas parece haber

| Nombre de la investigación | Objetivo | Área | Zona geográfica | Tamaño del corpus |
|---|---|---|---|---|
| Sketch Engine for bilingual Lexicography | Creación de diccionarios bilingües | Lexicografía | Brno, República Checa | 80 millones de palabras |
| European Union Language Resources in Sketch Engine | Análisis de texto, extracción de terminología y busqueda de textos para investigaciones | Sintaxis y Lexicografía | Unión europea | 850 millones de muestras en ingles |
| Corpus Based Extraction of hypernyms | Creación de diccionario terminológico | Lexicografía | República Checa | 9,757,005 palabras 27.389 documentos |
| Towards Automatic Finding of Word Sense Change in Time | Evaluación de procesos de obtención de corpus | Léxico | República Checa | 96.134.547 palabras (British National Corpus) |
| Software and Data for Corpus Pattern Analysis | Análisis de patrones para crear diccionarios | Lexicografía | República Checa | 100.000.000 palabras (British National Corpus) |
| Automatic generation of the Estonian Collocations Dictionary database | Se realizó un diccionario en estonio | Lexicografía | Tallinn, Estonia | 463 millones (Estonian National Corpus) |
| Semantic Word Sketches | Clasificación de palabras de forma automática | Léxico-semántico | Reino Unido | 352 textos (provenientes del Brown Corpus) |
| DIACRAN: a framework for diachronic analysis | Registrar el cambio de las palabras a lo largo del tiempo | Léxico | Reino Unido | 560 millones de palabras (provenientes del Corpus of Contemporary American English -COCA-) |
| Longest-commonest match | Desarrollo de un programa que localice colocaciones complejas de expresiones multipalabra | Lexicología | Reino Unido | 100.000.000 palabras (British National Corpus) |
| Effective corpus virtualization | Virtualización del corpus | Compilación de corpus | Reino Unido | Primeros 11 millones de palabras del British National Corpus y los primeros mil tokens del Brown corpus |
| Optimization of regular expressión evaluation within the manatee corpus management system | Optimización de la evaluación de expresiones regulares para la búsqueda en grandes corpus | Minería de texto | República Checa | 18.978.703 (czTenTen12); 27.894.538 (esTenTen12), 115.820.931 (ENClueWeb09) y 13.844.200 jpTenTen11 |
| arTenTen: Arabic corpus and word sketches | Creación de un corpus en arabe para apoyar la investigacón lingüística | Léxico | Reino Unido | 5.800.000.000 palabras |
| Hindi word Sketches | Creación de diccionario e investigación lingüística | Lexicografía | India and R. U. | 240 millones de palabras |
| Compatible Sketch grammars for compatible corpora | Experimento dirigido a crear una familia de corpora web de miles de millones de tokens | Lexicología | Distintas partes del mundo | Selección de muestras de corpus de francés, inglés, ruso, entre otros con una media de tamaño de 1 200 000 000 de palabras |
| Bilingual word Sketches: the traslate Button | Hacer traducciones efectivas para el corpus de Sketch Engine | Lexicografía | República Checa | Conjunto de corpus EUROPARL (provenientes de 22 lenguas oficiales de Europa) de entre 9 millones y 54 millones de palabras cada uno |
| Sketching the dependency relations of words in chinese | Establecer relaciones gramaticales a través del análisis sintáctico | Léxico | Taiwan | 2 100 000 000 palabras (zhTenTen11) |
| Intrinsic methods for comparison of corpora | Método de comparación cuantitativo entre generadores de corpus | Léxico | República Checa | 4.458.000.000 palabras (czTenTen12) y 2.607.000.000 palabras (Hector corpus) |
| Quantifying lexical usage: vocabulary pertaining to ecosystems and the environment | Descubrir colocaciones clave de términos ambientales seleccionados y establecer si tienden a usarse en contextos positivos, negativos o neutrales. | Léxico | Reino Unido | 1.500.000.000 de palabras (UKWaC) |
| Setting up for corpus lexicography | Creación y depuración de un corpus para realizar un diccionario en portugués | Lexicografía | República Checa | 3.500.000.000 millones de palabras |

TABLA 1. Corpus masivos para estudios de lingüística aplicada

marcado un camino bastante común en la planeación de sus diccionarios. Esta tendencia no cambió en dicha lengua y aún podemos constatarla en diccionarios tanto generales como especializados.

Se pensaría que algunos diccionarios terminológicos recurrirían a distintos tipos de información debido a lo complejo de sus definiciones, pero no es así. Incluso, como se ha demostrado en otro trabajo (Lázaro, 2011), esta tendencia parece reforzarse desde los años noventa y ya bien entrado el siglo XXI. Lo que contrasta con el avance tecnológico, pues se pensaría que las supresiones de elementos complementarios a la definición no deberían constituir un problema de diseño.

Las propuestas actuales para la elección de ejemplos no documentan una metodología de selección o creación. Lo más cercano ha sido que su búsqueda e identificación se ha llevado a veces como una tarea rigurosa que intenta colocar por lo menos uno de ellos en cada entrada, por ejemplo, en diccionarios como el COBUILD (Collins Birmingham University International Language) dirigido por Sinclair (1987) y el *Diccionario Planeta* (Marsá, 1989), o tal como lo intentó alguna vez Moliner en su célebre *Diccionario de uso del español* (Moliner, 2013).

En el ámbito teórico, los estudios que destacan la importancia del ejemplo tienen que ver sobre todo con diccionarios orientados a la traducción y la adquisición de una segunda lengua (Cowie, 1989; Drysdale, 1987; Fox, 1987; Humble, 1998; Laufer, 1992; Minaeva, 1992; Nesi, 1996; Paquot, 2008; Sinclair, 1987) y están basados, la mayoría de los casos, en sus funciones y el impacto en aspectos que tienen que ver con la educación. Meyer (2001), dentro de su concepción de *contextos ricos en conocimiento* (KRC), incluye los ejemplos y las definiciones como las representaciones implícitas comunes de una red conceptual.

El ejemplo en la tradición francesa ha sido un tema que ha captado de manera más evidente la atención de los lexicógrafos del siglo XX. Esta tradición ha sido resumida certeramente por Rey, quien nos dice que es un elemento derivado del análisis semántico en relación con la definición y que tiene una fuerte carga de ideología cultural (Rey, 1995).

En cuanto a su forma, en algunos casos se echa mano de colocaciones y estas se modifican mínimamente. Es decir, se conservan fragmentos textuales más amplios de lo que comúnmente encontraríamos como un ejemplo, lo que da lugar a un diccionario «contextualizador» (Mott y Mateo, 2009). En otros casos, hay ejemplos que parten de métodos computacionales, como la generación de concordancias y colocaciones complejas (Kilgarriff *et al.*, 2008); ejemplos confeccionados *ad hoc* (o inventados por el lexicógrafo) (Company, 2010), y en pocos casos se trata de fragmentos tomados del habla popular (Lara, 2008 y 2010).

Como podemos ver, en lexicografía hispánica no hay estudios de largo aliento que se dediquen exclusivamente al estudio del ejemplo. No es sino hasta los acercamientos de Jacinto García (2015) y Rojas Arregocés (2016) que tenemos obras completas en forma de libro o tesis dedicadas a esta categoría de datos. El primero hace una clasificación de los ejemplos atendiendo a su forma, origen o función y aborda un interesante y completísimo estudio que también toma en cuenta el punto de vista histórico y de autoridades, ya que este libro se desprende de su tesis doctoral *El principio de autoridad en los diccionarios generales del español (siglos XVIII-XX)*. La segunda hace un repaso sobre la forma y las funciones del ejemplo de acuerdo con lo documentado en obras de lexicografía general y, como los autores no hispanohablantes, lo orienta a la educación en su país natal, Colombia. De aquí se desprende una propuesta sobre cómo clasificar y elegir fragmentos textuales que se utilizarán como ejemplos en diccionarios escolares. Hay que mencionar que lo que la autora llama «programa» se refiere a pautas, no a un método específico ni a un programa informático, y por esta razón las conclusiones de su estudio desembocan en los argumentos sobre la importancia del ejemplo y las buenas prácticas en el proceso ejemplificativo, pero no hay una evaluación del método. Por esta razón, ambas obras son de corte teórico.

Nosotros, por otro lado, hemos decidido utilizar como punto de partida dichos estudios en la teoría, pero en la práctica nos inclinamos hacia un par de estudios para determinar la selección de un ejemplo en terminología. Por tanto, este será nuestro ámbito de trabajo y se justifica porque es un área más reducida y específica dentro del mundo de los diccionarios. Además, podemos encontrar que en este campo los esfuerzos han sido casi nulos, por lo que se desea aportar un estudio para el gremio terminológico.

Fuentes Morán y García Palacios (2002) hacen hincapié en la necesidad de los ejemplos en los diccionarios de especialidad y en el tratamiento profundo de la adecuación para elegir los más apropiados, como parte de la tipología de información que este tipo de diccionarios debe ofrecer (que aquí llamamos *categorías de datos*: registro, equivalentes, sinónimos, pronunciación, etc.). Del análisis que hacen los autores reproducimos un conjunto de características que el ejemplo debería cumplir y que está basado en lo dicho anteriormente por Gutiérrez Cuadrado (1999) y Drysdale (1987). A saber, que el ejemplo:

- Complementa la información de la definición.
- Muestra la palabra en su contexto.
- Distingue diferentes acepciones.
- Muestra colocaciones típicas.
- Indica registros y niveles estilísticos.
- Muestra diferentes cuestiones gramaticales.
- Incluye ciertas orientaciones ideológicas.
- Incluye información enciclopédica.
- Da información sobre costumbres y realidades específicas.

Aunado a estos criterios, tomamos el argumento de Meyer de que los ejemplos forman parte de una red conceptual y que, incluso, pueden encontrarse en un mismo contexto. A este tipo de contexto la autora lo ha llamado *knowledge rich context*. A saber:

> As a result of analyzing conceptual relations, a high-quality terminology project ultimately aims to illustrate the network of concepts underlying the terms of the domain. In traditional term banks and specialized dictionaries, the conceptual network is usually represented implicitly, through the definitions and examples provided to elucidate the meaning of a term. (Meyer, 2001, p. 280)

De los casos anteriores, tanto desde el punto lexicográfico como terminológico, rescatamos tres aspectos importantes: 1) los ejemplos pueden partir de fragmentos en los que el término muestra su aparición común, las concordancias; 2) algunos fragmentos en los que se inserta dicho término revelan su preferencia por ciertas combinaciones con otras piezas léxicas, como en el caso de las colocaciones, lo que nos muestra las palabras más indicadas para combinar con el término y diseñar ejemplificaciones, y 3) aun teniendo en cuenta los materiales que nos arrojan los corpus, tales como concordancias y colocaciones, es necesaria la afinación manual atendiendo a los criterios que, como especialista o hablante nativo de una lengua, tiene el lexicógrafo, tales como el conocimiento histórico, contextual, social, el habla espontánea no controlada o las modas lingüísticas. De esta forma podemos decir que los ejemplos conforman una esfera que complementa y amplía la información asociada a un término, y que actúan principalmente desde el concepto y no siempre desde la definición.

### 4.2 El paso previo: la *saturación semántica*

La *saturación semántica* es una noción teórica que intenta explicar el proceso ejemplificativo (Lázaro, 2016). Esto es importante porque, aun cuando los estudios sobre el ejemplo han profundizado bastante sobre su identificación, sus funciones, sus formas y sobre todo los criterios que pueden tomarse en cuenta para determinar si se trata de un «buen ejemplo» —de la tradición inglesa, *real examples* (Atkins y Rundell, 2008)—, su mecanismo de acción y el proceso por el que se forman no se han descrito. Tomando en cuenta esto, y bajo el argumento de que todo hablante puede emitir un ejemplo sobre una palabra con relativa facilidad, entonces inferimos que quizá sí existe un patrón para formar esta categoría de datos, pero no se identifica fácilmente, pues su sintaxis no está definida. Ahora bien, todo patrón se genera bajo algún mecanismo de formación. Si el mecanismo de formación no ha sido descrito, pero sus efectos son reconocibles, entonces es posible encontrarlo a través del análisis del conjunto de los efectos. Entiéndase *efectos*, aquí, como funciones y características. Todas las funciones del ejemplo, descritas y analizadas por los estudiosos que hemos citado, han permitido conformar la propuesta que describe el mecanismo de funcionamiento del ejemplo y, a través del análisis de ese mecanismo, proponemos una metodología plausible para su caracterización. La base del análisis es, por consiguiente, la relación existente entre los elementos que lo conforman: piezas léxicas, sus funciones y su intercambiabilidad. Por esto, se ha hecho necesaria la explicación sobre la manera en la que se relacionan término y concepto, término y definición, hasta dónde llega la definición y finalmente cuál es la función principal del ejemplo que la acompaña. Como vemos, aun sin contar con uno de los pilares de estudio de la lingüística, como es la sintaxis, puede llegarse a conclusiones aceptables a través de los efectos de las relaciones conceptuales, de eso que no vemos directamente en las estructuras del lenguaje. Si volvemos atrás una vez más, esto solo viene a comprobar que si el concepto es una abstracción de un segmento de la realidad, un proceso cognitivo, no podía esperarse que uno de sus descriptores, el ejemplo, fuese un proceso alejado de ese mecanismo. La propuesta es la citada *saturación semántica*.

### 4.3 La *densidad lexicométrica*: un método para identificar ejemplos en corpus masivos

En un experimento llevado a cabo con el paquete de herramientas proporcionado por Sketch Engine, diseñamos una metodología formal para crear un algoritmo basado en las características teóricas que debe cumplir un ejemplo. Elegidos quince términos de tres áreas temáticas distintas (sexualidad, finanzas e informática), se aplicó el siguiente flujo de trabajo: 1) extraer todas las concordancias en las que aparece un término; 2) a partir de esas concordancias, hacer un análisis con Word Sketch, esto es, determinar las relaciones sintácticas del término con las palabras que lo rodean a fin de conocer cuáles son los verbos que se le asocian; 3) hacer una búsqueda de concordancia compleja en la que la *consulta* fuese el término elegido y el *contexto*, el verbo con la asociación más estrecha, esto es, un fragmento textual en el que término y verbo aparecieran en el mismo contexto, y 4) llevar a cabo una extracción de colocaciones complejas a partir de los fragmentos que cumpliesen las restricciones anteriores con el fin de determinar qué conjugación de ese verbo asociado es más productiva para cumplir los criterios de un ejemplo.[7] La tabla 2 muestra los términos de sexualidad y sus equivalentes. Se hizo lo mismo con las dos áreas restantes.

| Término | Equivalentes |
|---|---|
| Bisexualidad | Preferencia bisexual |
| Coito | |
| Diafragma | |
| Erotismo | |
| Género | |
| Heterosexualidad | |
| Homosexualidad | |
| Identidad de género | Identidad sexual |
| Intersexualidad | |
| Preferencia sexual | Orientación sexual |
| Sexo | |
| Transexualidad | |
| Travestismo | |
| Sexualidad | |
| Papel genérico estereotipado | Rol sexual, papel genérico |

TABLA 2. Términos y equivalentes en el área de sexualidad

En un primer acercamiento aplicamos este flujo de trabajo a cinco términos del área de la sexualidad dentro de un corpus no muy grande (aproximadamente, 1 millón de palabras), el Corpus de las Sexualidades de México[8]: desafortunadamente, el resultado fue infructífero, pues el número de fragmentos que cumplieron todas las restricciones fue igual a cero. Sin embargo, durante el proceso fue posible observar que si bien el número de fragmentos iba reduciéndose, estos tenían ciertas características que los acercaban a un ejemplo teórico, por lo que la solución que se encontró fue hacer búsquedas en corpus más grandes. A continuación los describimos.

*La Jornada* es un diario mexicano de circulación nacional, que apareció en 1996. El corpus que creamos está formado por todo el contenido del sitio *http://www.jornada.unam.mx* hasta el 30 de mayo de 2014, incluyendo los comentarios de los usuarios, ya que fue extraído automáticamente de su página de Internet. Básicamente es un corpus periodístico, sincrónico, de registro culto y con inclinación hacia la política, la economía y la cultura. Este corpus contiene 382 millones de palabras (382.119.353), distribuidas aproximadamente en 15 millones de frases.

Corpus Wiki es el corpus formado con todos los artículos de la *Wikipedia* en español (versión de septiembre del 2014). Se trata de un corpus de lengua general, enciclopédico, de registros varios y sincrónico. Está formado por 392 millones de palabras (392.530.981), distribuidas aproximadamente en 16 millones de frases.

Corpus esTenTen (5Go) es una versión reducida del original esTenTen, de Sketch Engine (Kilgarriff y Renau, 2013), en la que solo utilizamos los primeros 5 gigabytes de texto. Es un corpus de lengua general que agrupa todas las variantes del español. Esta muestra tiene un tamaño de más de 2.000 millones de palabras (2.443.447.212), distribuidas aproximadamente en 78 millones de frases.

Todos los corpus fueron convertidos a texto plano, limpiados, recodificados en UTF-8, divididos en frases usando parte del sistema Cortex (Torres *et al.*, 2002) y guardados en una versión etiquetada.

Este análisis, a pequeña escala (con unos miles de frases), reveló que el mecanismo por el cual opera un ejemplo es el de la ampliación de los rasgos de un concepto y la reducción de los contextos en los que aparece. Pero, como puede inferirse, llevar a cabo el flujo de trabajo que hemos descrito y aplicarlo a corpus de millones de frases se convertía en una tarea titánica. Por lo anterior, se diseñó un programa capaz de ejecutar de manera automática ese trabajo: Genex: Générateur d'Examples (Lázaro, 2015). Este sistema funciona básicamente midiendo el número de palabras que hay en la definición asociada a un término, calculando la información mutua que hay entre ellas y luego eligiendo los fragmentos más cortos, divididos como frases en los corpus presentados, que cumplan con todos los requerimientos que hemos presentado al principio de este apartado.

En otras palabras, la *densidad lexicométrica* es un producto directo de la noción de *saturación semántica*. Para medir la *densidad lexicométrica*, es necesario tener en cuenta que la definición terminográfica guarda una relación estrecha con el ejemplo. Las piezas léxicas de la primera deberían aparecer idealmente en el segundo, aunque no en un estricto orden sintáctico. La aparición de los elementos de la definición en el ejemplo at inación de las palabras de una definición en una nueva estructura no definitoria permite que el contexto de activación (Kuguel, 2007) del término siga operando y concede al fragmento una pertenencia a un campo semántico determinado, que en este caso será un área de especialidad. A esto le hemos llamado *contexto nominal* (Lázaro, 2015, p. 145).

Por otro lado, quitar a la definición los verbos definitorios ha conllevado que el nuevo fragmento carezca de estructura. Para solventar esto, se ha decidido colocar en esta nueva estructura el verbo semánticamente más cercano al término. Dicho verbo puede existir *a priori* en la definición o puede ser producto de una búsqueda hecha con la medida de *información mutua*. Todo esto sirve para dar estabilidad a la nueva estructura: a esto le hemos llamado *contexto verbal* (Lázaro, 2015, p. 146).

Así, la *densidad lexicométrica* calcula qué tanta información tiene un fragmento textual y qué tanta cercanía semántica tiene con respecto a otro fragmento (que en este caso será la definición del término), de tal manera que puede complementarlo en el plano conceptual. La *densidad lexicométrica* puede considerarse como el producto del valor coseno de una frase con respecto a otra por el inverso del logaritmo de su longitud. Gráficamente, la fórmula para determinarla es la siguiente:

$$densidad\ lexicométrica_i = \cos(frase_i, q) \times \frac{1}{\log |frase_i|}$$

## 5. Resultados

La evaluación del sistema se hizo a través de una encuesta que constaba de 15 ejemplos por cada término (5 de cada corpus). Es decir, una muestra de 225 ítems cuya función era determinar qué tanto el sistema se acercaba a lo que un hablante de español considera como un buen ejemplo. Los sujetos de evaluación, 50 hablantes nativos de español, tenían entre 18 y 70 años y eran 31 mujeres y 19 varones. Su rango de educación estaba entre el bachillerato (6), la universidad (31) y el posgrado (13). Es decir, diastráticamente pertenecían al registro culto. Su función fue elegir, entre los 5 candidatos que generó el sistema, el que ellos consideraban como el mejor ejemplo para cada término. Tuvieron 3 secciones, una por cada corpus. Los ejemplos en dichas secciones estaban acomodados en orden descendente de acuerdo con la puntuación determinada por la *densidad lexicométrica*, esto es, los más cercanos semánticamente a la definición y más cortos según su longitud de palabras estaban en primer lugar. Los resultados obtenidos fueron los que se muestran en la tabla 3.

Los datos arrojados muestran la correlación existente entre lo que dice el sistema y lo que prefieren los hablantes. La evaluación puede ir de –1 a +1. Cuando el resultado se acerca a +1, quiere decir que el humano y el sistema están de acuerdo en que la selección de un candidato a ejemplo es ideal. Cuanto más alejado esté, es decir, cuando se acerca a –1, es que el sistema no logró ordenar precisamente los candidatos. Una evaluación cercana a 0 significa que no hay acuerdo ni desacuerdo. La tabla nos muestra los resultados para cada término en cada corpus, así como el coeficiente por área.

Es de notar que en el corpus Jornada el sistema tuvo serias dificultades para encontrar buenos candidatos a ejemplo. Esto puede explicarse por dos razones básicas: 1) se trata de un corpus periodístico que tiene muy acotadas sus áreas y su lenguaje se acerca mucho a lo especializado, y 2) las temáticas centrales de este diario son las finanzas, la economía y la política, y los resultados del área de finanzas son positivos, mientras que los coeficientes de las áreas de sexualidad e informática caen drásticamente.

Por otra parte, el corpus Wikipedia fue donde el sistema pudo trabajar mejor y obtener mejores candidatos a ejemplo de acuerdo con la evaluación humana. Incluso es notable el hecho de que no se dio una correlación lineal ($r = 0$), como sí ocurrió en esTenTen, o una correlación negativa ($-1 < |r| < 0$), como tiene el amplio rango de Jornada.

Finalmente, en el corpus esTenTen, Genex tuvo algunos resultados poco alentadores, pero también encontró información suficiente para generar algunos ejemplos, sobre todo en el área de informática. Suponemos que su bajo rendimiento se debe al hecho de que es una muestra pequeña (5 gigabytes) de un corpus mucho más grande (16 gigabytes), lo que pudo ocasionar que la muestra extraída fuera poco representativa, ya que no se controló su variedad ni su equilibrio.

| | | Jornada | | Wikipedia | | EsTenTen | |
|---|---|---|---|---|---|---|---|
| | | Término | Área | Término | Área | Término | Área |
| **sexualidad** | sexualidad | -0,2 | -0,2 | 0,8 | 0,04 | 0,9 | 0 |
| | sexo | -0,3 | | 0,5 | | -0,7 | |
| | aborto | 0,7 | | -0,5 | | 0 | |
| | matrimonio | -0,8 | | -0,8 | | 0,2 | |
| | embarazo | -0,4 | | 0,2 | | 0,2 | |
| **finanzas** | banco | -0,1 | 0,24 | 0,5 | 0,16 | -0,1 | 0,1 |
| | crédito | 0,3 | | 1 | | -0,8 | |
| | nómina | 0,5 | | -0,6 | | 0,7 | |
| | préstamo | 0,3 | | 0,2 | | 0,2 | |
| | pensión | 0,2 | | -0,3 | | 0,5 | |
| **informática** | virus | 0,6 | -0,14 | 0,1 | 0,26 | 0 | 0,08 |
| | computadora | -0,1 | | -0,3 | | 0,3 | |
| | programa | -0,6 | | 0,9 | | -0,2 | |
| | teclado | 0,2 | | 0,2 | | 0,7 | |
| | informática | -0,8 | | 0,4 | | -0,4 | |
| | **Promedio** | **-0,033 333 3** | | **0,153 333 33** | | **0,1** | |

TABLA 3. Resultados por coeficiente de correlación de Pearson

## 6. Conclusiones

El sistema Genex ha demostrado que es posible obtener ejemplos automáticamente con pocas reglas lingüísticas y sobre todo a través de la delimitación de contextos. Esto comprueba que un ejemplo no tiene una estructura sintáctica fija, sino que es el resultado de acotaciones que tienen que ver con criterios semánticos y de asociación de palabras.

El funcionamiento básico del sistema trata de imitar el proceso mental que llevamos a cabo como humanos para seleccionar un ejemplo. Es decir, elige de un conjunto ilimitado de posibles contextos en los que aparece un término, aquel que guarde una relación semántica estrecha con la definición de ese mismo término. Tiene como criterios de selección el conjunto de palabras que conforman una definición determinada. Un ejemplo elegido por un humano parte de su conocimiento previo del concepto del término que desea ejemplificar. En GENEX, este conocimiento es imitado a través de un repositorio organizado de dichos conceptos que se realizan formalmente en definiciones terminográficas. Por tanto, un candidato a ejemplo es aquel contexto que, teniendo como núcleo un término, a través de un conjunto de restricciones, proyecta rasgos distintos a los de una definición y la complementa.

En esta investigación observamos que el corpus Jornada impedía que el sistema obtuviera resultados positivos ($r|< 0$), lo que comprueba que un corpus especializado, que a su vez tendrá sobre todo contextos especializados, pocas veces o ninguna puede funcionar para diseñar un ejemplo. Los otros dos corpus eran compatibles con el sistema debido a su naturaleza más general. De esta manera, podemos argüir que el ejemplo es el resultado de llevar un término a un contexto general manteniendo las características que lo hacen perteneciente a un dominio especializado. El ejemplo, así, sirve para aclarar la opacidad comprensiva que puede crear una definición terminológica.

En términos generales, podemos esbozar que existe en la actualidad un cambio de paradigma en la investigación lingüística. Ya no se estudia el lenguaje conforme a conclusiones sacadas a partir del conocimiento lingüístico del especialista y sostenidas por algunos cientos de muestras (en el mejor de los casos), sino que se parte del tipo de teorizaciones mostradas en este estudio de caso y se crean modelos y algoritmos para comprobarlos en la totalidad de un corpus masivo. Si ahora, en el 2019, decidiésemos mostrar pruebas fehacientes sobre un postulado lingüístico, sería inocente pensar que cinco, ocho o diez ejemplos en un artículo serían suficientes. Incluso sería difícil pensar en una tesis que proponga doscientos casos para poner en marcha el análisis de un paradigma en la lingüística actual.

Hay que notar también, sin embargo, que efectivamente algunos casos son susceptibles de análisis con una presentación mínima de casos de estudio o muestras, como los diacrónicos. Pero lo que hemos intentado mostrar aquí es que esa tendencia definitivamente va a la baja y es muy probable que en poco tiempo se reserve solo a cierto tipo de casos y que la lingüística computacional y la de corpus sean elementos indispensables incluso en los estudios históricos.

## Bibliografía

Atkins, B. T. Sue y Rundell, Michael. (2008). *The Oxford guide to Practical Lexicography*. Oxford: Oxford University Press.

Cabré, M. Teresa (1993). *La terminología: Teoría, metodología, aplicaciones*. Barcelona: Empúries.

Carnie, Andrew (2013). *Syntax: A generative introduction*. West Sussex: Wiley-Blackwell.

Company, Concepción. (coord.) (2010). *Diccionario de Mexicanismos*. México D.F.: Siglo XXI editores/Academia Mexicana de la Lengua.

Cowie, Anthony P. (1989). «The language of examples in English learners' dictionaries». A: James, Gregory (ed.). *Lexicographers and their works*. Exeter: University of Exeter (Exeter Linguistic Studies), p. 55-65.

Cunha, Iria da; San Juan, Eric; Torres-Moreno, Juan-Manuel; Cabré, M. Teresa; Sierra, Gerardo (2012). «A symbolic approach for automatic detection of nuclearity and rhetorical relations among intra-sentence discourse segments in Spanish». A: *Computational linguistics and intelligent text processing* (Lecture Notes in Computer Science, LNCS), p. 462-474.

Drysdale, Patrick (1987). «The role of examples in a learner's dictionary». A: Cowie, Anthony P. (ed.). *The dictionary and the language learner: Papers from the EURALEX Seminar at the University of Leeds, 1-3 April 1985*. Tubinga: Max Niemeyer (Lexicographica. Series Maior; 17), p. 213-223.

Fox, Gwyneth (1987). «The case for examples». A: Sinclair, John (ed.). *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. Londres: Collins ELT, p. 137-149.
Fuentes Morán, M. Teresa; García Palacios, Joaquín (2002). «Los ejemplos en el diccionario de especialidad». A: García Palacios, Joaquín; Fuentes Morán, M. Teresa (ed.). *Texto, terminología y traducción*. Salamanca: Almar, p. 75-98.
Fuertes-Olivera, Pedro A.; Bergenholtz, Henning; Nielsen, Sandro; y Amo, Marta (2012). «Classification in Lexicography: The Concept of Collocation in the Accounting Dictionaries». *Lexicographica*, 28 (1), 293-308.
Gutiérrez Cuadrado, Juan (1999). «Notas a propósito de la ejemplificación y la sinonimia en los diccionarios para extranjeros». A: Vila Rubio, M. Neus; Calero Fernández, M. Ángeles; Mateu, Rosa; Casanovas Català, Montserrat; Orduña López, José Luis (coord.). *Así son los diccionarios*. Lleida: Universitat de Lleida, p. 77-98.
Humble, Philippe (1998). «The use of authentic, made-up, and controlled examples in foreign language dictionaries». A: Fontenelle, Thierry; Hiligsmann, Philippe; Michiels, Archibald; Moulin, André; Theissen, Siegfried (ed.). *EURALEX'98 Proceedings*. Lieja: Université de Liège, p. 593-599.
Jacinto García, Eduardo (2015). *Forma y función del diccionario: Hacia una teoría general del ejemplo lexicográfico*. Jaén: Universidad de Jaén, p. 146.
Kilgarriff, Adam; Husák, Miloš; McAdam, Katy; Rundell, Michael; Rychlý, Pavel (2008). «GDEX: Automatically finding good dictionary examples in a corpus» A: *Proceedings of the 13th EURALEX International Congress*. Barcelona: IULA, p. 425-432.
Kilgarriff, Adam; Renau, Irene (2013). «esTenTen, a vast web corpus of Peninsular and American Spanish». *Procedia: Social and Behavioral Sciences*, vol. 95, p. 12-19.
Kuguel, Inés (2007). «La activación del significado especializado». A: Lorente, Mercè; Estopà, Rosa; Freixa, Judit; Martí, Jaume; Tebé, Carles (ed.). *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví*. Barcelona: IULA, p. 337-354.
Lara, Luis Fernando (2008). *Diccionario del español usual en México*. 2.ª ed. México: El Colegio de México.
Lara, Luis Fernando (2010). *Diccionario del español de México*. 1.ª ed. México: El Colegio de México.
Laufer, Batia (1992). «Corpus-based versus lexicographer examples in comprehension and production of new words». A: Tommola, Hannu; Varantola, Krista. (ed.). *EURALEX'92 Proceedings*. Tampere: University of Tampere, p. 71-76.
Lázaro, Jorge (2011). *Análisis de relevancia cuantitativa y cualitativa de ejemplos y contextos de uso en definiciones de términos referidos a sexualidad*. Trabajo de fin de máster. Barcelona: IULA.
Lázaro, Jorge (2015). *El ejemplo en terminología: Caracterización y extracción automática*. Tesis doctoral. Barcelona: IULA.
Lázaro, Jorge (2016). «Sobre la noción de saturación semántica en terminología». *Debate Terminológico*, 15, p. 66-81.
López-Escobedo, Fernanda; Solórzano-Soto, Julián; Sierra, Gerardo (2016). «Analysis of intertextual distances using multidimensional scaling in the context of authorship attribution». *Journal of Quantitative Linguistics*, vol. 23, núm. 2, p. 154-176.
Marsá, Francisco. (1982). *Diccionario Planeta de la lengua española usual*. Indiana: Planeta Publishing Corporation.
McEnery, Tony; Hardie, Andrew (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
Mayer-Schönberger, Viktor; Cukier, Kenneth (2013). *Big data: La revolución de los datos masivos*. Madrid: Turner.
Meyer, Ingrid (2001). «Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework». A: Bourigault, Didier; Jacquemin, Christian; L'Homme, Marie-Claude. *Recent advances in computational terminology*. Amsterdam: John Benjamins, p. 279-302.
Minaeva, Ludmila (1992). «Dictionary examples: friends or foes?». A: Tommola, Hannu; Varantola, Krista (ed.). *EURALEX'92 Proceedings*. Tampere: University of Tampere, p. 77-80.
Moliner, María (2013). *Diccionario de uso del español*. 3.ª ed. Madrid: Gredos.
Mott, Brian; Mateo, Marta (2009). *Diccionario-guía de traducción español-inglés, inglés-español*. Barcelona: Universitat de Barcelona.
Nesi, Hilary (1996). «The role of illustrative examples in productive dictionary use». *Dictionaries: Journal of the Dictionary Society of North America*, vol. 17, núm. 1, p. 198-206.
Paquot, Magali (2008). «Exemplification in learner writing: a cross-linguistic perspective». A: Meunier, Fanny; Granger, Sylviane (ed.). *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins, p. 101-119.

Rey, Alain. (1995). «Du discours au discours par l'usage : pour une problématique de l'exemple». A: *Langue française, n°106. L'exemple dans le dictionnaire de langue Histoire, typologie, problématique, sous la direction de Alise Lehmann*. pp. 95-120.
Rojas Arregocés, Eufrocina (2016). *La ejemplificación en los diccionarios escolares en Colombia: Una revisión en tres diccionarios para bachillerato*. Tesis doctoral. Barcelona: IULA.
Sardinha, Tony Berber (2000). «Lingüística de corpus: histórico e problemática». *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, vol. 16, núm. 2, p. 323-367.
Sierra, Gerardo (2008). «Diseño de corpus textuales para fines lingüísticos». A: *IX Encuentro Internacional de Lingüística en el Noroeste*. Vol. 2. Sonora: UNISON, p. 445-462.
Sierra, Gerardo; Alarcón, Rodrigo; Aguilar, César; Bach, Carme (2008). «Definitional verbal patterns for semantic relation extraction». *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication* [Amsterdam: John Benjamins], vol. 14, núm 1, p. 74-98.
Sinclair, John (1987). *Collins Birmingham University International Language Database*. Londres: Collins.
Torres-Moreno, Juan-Manuel (2014). *Automatic text summarization*. Londres: John Wiley & Sons.
Torruella, Joan; Llisterri, Joaquim (1999). «Diseño de corpus textuales y orales». A: Blecua, José Manuel; Clavería, Gloria; Sánchez, Carlos; Torruella, Joan (ed.). *Filología e informática: Nuevas tecnologías en los estudios filológicos*. Barcelona: Milenio, p. 45-77.
Torres-Moreno, Juan Manuel; Velázquez-Morales, Patricia; y Meunier, Jean-Guy (2002). «Condensés de textes par des méthodes numériques». *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*. Saint-Malo: Lexícometríca. Revue électroníque. 1-12.

**Notas**

1. Un caso interesante sobre suplantación de identidad para cometer un delito de fraude bancario es el que puede leerse en una sentencia de la Sala de lo Penal del Tribunal Supremo de Madrid, del año 2007, que dio pie a la formulación de varias leyes que ahora forman parte del Código Penal de España. La sentencia completa puede leerse en *http://web.icam.es/bucket/Faustino%20Gud%C3%ADn%20-%20Nuevos%20delitos%20inform%C3%A1ticos.pdf*.
2. Pensemos aquí en las *colocaciones*, estas estructuras recurrentes a medio camino de la lexicalización que no son posibles de ver en corpus pequeños. Un estudio interesante sobre la identificación de colocaciones en lexicografía, aplicado a veintitrés diccionarios a la vez, podemos encontrarlo en el estudio de Fuertes-Olivera *et al.* (2012).
3. *http://corpus.rae.es/creanet.html*.
4. *http://corpus.rae.es/cordenet.html*.
5. Datos de un volcado de septiembre del 2014.
6. Esta tabla se creó a partir de los datos de Sketch Engine. Todos ellos pueden encontrarse en *https://www.sketchengine.eu*.
7. Cabe mencionar que esta restricción es importante, ya que hay casos en los que, para ejemplificar un término, utilizamos preferentemente ciertos tiempos verbales debido a la naturaleza misma de la palabra. Es decir, es más probable que se hable de la infección de un *virus* una vez que esta ha sucedido («El ordenador fue infectado por un *virus*») que en otros tiempos que romperían la lógica de cómo nos expresamos a propósito de ese término (** «El ordenador será infectado por un *virus*», * «Infectaré mi ordenador con ese *virus*»).
8. *www.corpus.unam.mx/csmx/*