La lingüística de corpus aplicada al desarrollo de la competencia tecnológica en los estudios de traducción e interpretación y la enseñanza de segundas lenguas

STUDIEN ZUR ROMANISCHEN
SPRACHWISSENSCHAFT UND
INTERKULTURELLEN KOMMUNIKATION

Herausgegeben von Gerd Wotjak

BAND 127

PETER LANG

Miriam Seghiri

La lingüística de corpus aplicada al desarrollo de la competencia tecnológica en los estudios de traducción e interpretación y la enseñanza de segundas lenguas

PETER LANG

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Gedruckt auf alterungsbeständigem, säurefreiem Papier.
Druck und Bindung: CPI books GmbH, Leck

ISSN 1436-1914
ISBN 978-3-631-74122-1 (Print)
E-ISBN 978-3-631-76945-4 (E-PDF)
E-ISBN 978-3-631-76946-1 (EPUB)
E-ISBN 978-3-631-76947-8 (MOBI)
DOI 10.3726/b14734

Peter Lang – Berlin · Bern · Bruxelles ·
New York · Oxford · Warszawa · Wien

Diese Publikation wurde begutachtet.

www.peterlang.com

Comité científico y revisor

1. Elena Bandín Fuertes. Universidad de León
2. Soledad Díaz Alarcón. Universidad de Córdoba
3. José María Díaz Lage. Universidad Internacional de La Rioja
4. Diana Esteba Ramos. Universidad de Málaga
5. Estefanía Flores Acuña. Universidad Pablo de Olavide
6. Ángeles García Calderón. Universidad de Córdoba
7. Dunia Hourani Matín. Universität Leipzig (Alemania)
8. Andrés Jiménez Fernández. Universidad de Málaga
9. Carmen B. Macías Corredera. Universidad de Murcia
10. Marie Hélène Maux. Université de Estrasburgo (Francia)
11. Esteban T. Montoro del Arco. Universidad de Granada
12. Mazal Oaknin. University College de Londres (Reino Unido)
13. Leonor Pérez Ruiz. Universidad de Valladolid
14. Mercedes Querol Julián. Universidad internacional de La Rioja
15. María Recuenco Peñalver. Universidad de Ciudad del Cabo (Sudáfrica)
16. María del Pilar Rodríguez Reina. Universidad Pablo de Olavide
17. Beatriz Rubio Martínez. University College Dublin (Irlanda)
18. Aurora Ruiz Mezcua. Universidad de Córdoba
19. Daniel Moisés Sáez Rivera. Universidad Complutense de Madrid
20. Isabel María Sánchez Arriaza. Universidad de Alcalá
21. Inmaculada Clotilde Santos Díaz. Universidad de Málaga
22. Juan Antonio Solís Becerra. Universidad de Murcia
23. Hanane Benali Taouis. The American University of the Middle East (Kuwait)
24. Milagros Torrado Cespon. Universidad internacional de La Rioja
25. María Azahara Veroz González. Universidad de Córdoba
26. Francisco Javier Vigier Moreno. Universidad Pablo de Olavide

Índice

M.ª Cristina Toledo Báez

Departamento de Traducción e Interpretación, Filología Francesa,
Estudios Semíticos y Documentación
Universidad de Córdoba
cristina.toledo@uco.es

Sketch Engine en traducción científico-técnica (francés-español): creación y explotación del corpus *ad hoc* comparable *GeneCorp*

Abstract: Lack of appropriate terminological and lexicographical resources is one of the major obstacles that translators face when dealing with specialised translation. Virtual corpora can compensate for this shortage of resources as they allow translators to create and exploit their own terminological material. The task of manually compiling corpora was arduous and tedious years ago, but thanks to Internet, nowadays it is possible to compile corpora semi-automatically. In addition, it has affected the way in which we conceive of corpora today and the web is even seen a corpus supermarket. Both innovations —the semi-automatic compilation of corpora and the conception of web as a supermarket— are part of our study thanks to Sketch Engine, a corpus manager and text analysis software, and its corpus-building tool WebBootCat. The aim of our study is to use both WebBootCat and Sketch Engine to, first, semi-automatically compile and, second, exploit *GeneCorp*, our bilingual (French-Spanish), comparable virtual corpus on Genetics. The results of our study show that the combination of both Sketch Engine and our *GeneCorp* corpus is very convenient as it brings four useful functionalities for translators: 1) the terminological functionality to obtain monolexical and polilexical terminological unit lists; 2) concordance analysis functionality to analyse both collocation behaviour and phraseological pattern of terms; 3) monolingual and bilingual syntactic, grammatical, and collocation behaviour analysis functionality; 4) distributional thesaurus functionality to clarify conceptual doubts. The combination of virtual bilingual corpora and Sketch Engine helps translators find very quickly no-cost high-quality terminological, phraseological, and thematic resources. Its use is proven to be beneficial for translation lecturers and professional translators and/or interpreters.

Keywords: bilingual virtual corpora, Sketch Engine, WebBootCat, web as corpus supermarket, semi-automatic corpora compilation.

Resumen: A la hora de enfrentarse a una traducción especializada, el traductor se encuentra con una falta de recursos lexicográficos y terminológicos apropiados, de ahí que el corpus virtual emerja como solución que le permite crearse su propio material de consulta.

Si hace años compilar corpus de forma manual constituía una tarea tediosa, en la actualidad Internet ha traído consigo la posibilidad de, por un lado, compilar corpus de forma semiautomática, y, por otro lado, considerar la red como supermercado de corpus que le permita descargar corpus de forma más eficiente. En el presente estudio hacemos uso de ambos elementos gracias a la herramienta de corpus en línea Sketch Engine y a su compilador WebBootCat. El objetivo principal de nuestro trabajo es emplear WebBoot Cat y Sketch Engine tanto para compilar de forma semiautomática como para explotar *GeneCorp*, un corpus *ad hoc* comparable bilingüe (francés-español) sobre genética. En los resultados de nuestro trabajo hemos demostrado cómo Sketch Engine, junto con nuestro corpus *GeneCorp*, constituye una herramienta de interés al combinar cuatro funcionalidades sumamente útiles para la traducción: 1) terminológica, que permite obtener listas bilingües de unidades terminológicas monoléxicas y poliléxicas; 2) de análisis de concordancias, con la cual se analiza el patrón colocacional y fraseológico de un determinado término; 3) de análisis de patrones sintácticos, gramaticales y colocaciones, patrones que pueden ser tanto monolingües como bilingües; 4) de creación de tesauros distribucionales, útil para aclarar dudas conceptuales. La combinación de corpus *ad hoc* bilingües y Sketch Engine ayuda a obtener recursos terminológicos, fraseológicos y temáticos de calidad, a coste cero y con suma rapidez, con lo que su uso tanto en el aula de Traducción como en la labor profesional del traductor y del intérprete se ha demostrado que es altamente recomendable.

Palabras clave: corpus *ad hoc* comparable bilingüe, Sketch Engine, WebBootCat, web como supermercado de corpus, compilación semiautomática de corpus.

1. Introducción

A pesar de los avances en lexicografía y terminografía y de la cantidad ingente de diccionarios especializados, bases de datos o glosarios, los traductores se siguen encontrando con una dificultad de gran calibre a la hora de traducir textos especializados: la falta de recursos lexicográficos y terminológicos que respondan a sus necesidades y que les permita adquirir un conocimiento experto. La carencia de recursos puede deberse a la especialización del texto, a la novedad temática, al par de lenguas de trabajo o a una combinación de cualquier de los factores anteriores. Ante la escasez de recursos, el traductor se ve obligado a crear sus propios materiales y es en ese punto cuando se hace imprescindible el uso de *corpus virtuales* o *ad hoc*, es decir, colecciones de textos extraídos de Internet y compilados con el objetivo de proporcionar información temática y lingüística de un determinado campo y para ser utilizados en la realización de una tarea concreta (Sánchez Gijón, 2009).

Son numerosos los autores que informan de los beneficios y de las implicaciones didácticas del uso del corpus virtual (Bowker y Pearson, 2002, Zanettin et al., 2003; Corpas Pastor, 2008; Castillo Rodríguez, 2014; Corpas Pastor y Seghiri,

2016; Seghiri, 2017; Toledo Báez y Martínez Lorente, 2018), tanto si se trata de *corpus comparables*, esto es, documentos redactados en lengua original que han sido compilados atendiendo a unos mismos criterios de selección (Corpas Pastor, 2001), como *corpus paralelos*, los cuales están formados "por una serie de textos en la lengua de origen junto con sus traducciones en una (o varias) lengua(s) meta" (Corpas Pastor, 2001, pág. 158). En efecto, la metodología CULT: *Corpus Use and Learning to Translate* (Beeby et al., 2009), desde su creación en 1997, ha permitido, como señala Sánchez Ramos (2017a), el desarrollo de las competencias que un traductor debe poseer, como son, entre otras, las que atañen a la adquisición de terminología y al manejo de distintas herramientas documentales, entre las que destaca el manejo de gestores de concordancias.

A pesar de las claras ventajas que ofrecen, compilar corpus virtuales resultaba una tarea tediosa años atrás cuando la compilación tenía que ser exclusivamente manual. Entre sus múltiples beneficios, tanto Internet como la imbricación de Lingüística de Corpus y de Lingüística Computacional han traído consigo nuevas formas de compilar y acceder a corpus. Son dos los principales cambios que han propiciado: por un lado, el uso de Internet y su relación con los corpus, empleando incluso la red como corpus (Gatto, 2013) o como supermercado de corpus (Bernardini et al., 2006) y, por otro, la automatización o semiautomatización de la compilación de corpus (Esplà-Gomis y Forcada, 2010) y de sus correspondientes fases, a saber, búsqueda y acceso de la documentación, descarga de datos, formato y almacenamiento (Seghiri, 2011, 2017).

En el presente estudio combinaremos ambos avances al hacer uso, de una parte, de la web como supermercado de corpus y, de otra, de la compilación semiautomática de corpus gracias a la herramienta de análisis de corpus en línea Sketch Engine[1] (Kilgarriff et al., 2004, 2014), la cual integra el compilador de corpus WebBootCat (Baroni et al., 2006). El objetivo principal de nuestro trabajo es emplear Sketch Engine para compilar y explotar *GeneCorp*, un corpus *ad hoc* comparable bilingüe (francés-español) sobre genética, y facilitar así la fase de documentación temática y terminológica previa a la traducción de un texto científico sobre genética, que se abordará en la asignatura Traducción científica y técnica de la lengua B (francés) del Grado en Traducción e Interpretación de la Universidad de Córdoba. En aras de alcanzar este objetivo principal, nos hemos planteado los siguientes objetivos secundarios:

1 Sketch Engine se encuentra disponible en la URL <https://www.sketchengine.co.uk/>. Es preciso registrarse para poder acceder a todas sus funcionalidades, incluyendo WebBootCat.

1. Aproximarnos desde el punto de vista teórico a, por un lado, la web para corpus y la web como corpus y, por otro, a la compilación automática y semiautomática de corpus.
2. Presentar la herramienta Sketch Engine y detallar sus distintas funcionalidades, en particular WebBootCat.
3. En función del encargo de traducción asignado, compilar de forma semiautomática con WebBootCat, un corpus *ad hoc* comparable bilingüe (francés-español) sobre genética que hemos denominado *GeneCorp*.
4. Explotar el corpus *GeneCorp* a través de Sketch Engine y de sus distintas funcionalidades para facilitar la labor terminológica y la documentación temática del traductor.

No podemos obviar que son varios los trabajos anteriores que han empleado Sketch Engine en el aula de traducción científico-técnica, ya sea para la combinación lingüística inglés-español (López Rodríguez y Buendía Castro, 2011; López Rodríguez, 2016; Sánchez Ramos, 2017a) o inglés-francés (Rossi et al., 2016; Borel Tagne y Loock, 2017). La particularidad de nuestro trabajo es doble: 1) nos centraremos en el par de lenguas francés-español, escasamente estudiada en lo que al uso de Sketch Engine conlleva, y 2) nuestro enfoque no estará centrado únicamente en el aula, sino que también puede extrapolarse a la traducción profesional e incluso a la interpretación.

2. La red y la lingüística de corpus: implicaciones en la compilación y en la explotación de corpus

En este apartado nos centraremos en los dos principales cambios que la unión de Internet, la Lingüística de Corpus y la Lingüística Computacional han traído consigo: la propia concepción de corpus, llegando a considerar la red como un corpus en sí mismo, y la automatización o semiautomatización del proceso de compilación de corpus.

2.1. Web para corpus y web como corpus

La expansión de Internet ha traído consigo nuevas formas de organizar y obtener la información que se traducen también en nuevas formas de compilar corpus. Esta influencia de Internet ha hecho que los estudiosos de la Lingüística de corpus se planteen cómo usar la red. En este sentido, Fletcher (2007) plantea dos enfoques para usar Internet en la investigación en Lingüística de corpus:

La web para corpus o web para recopilar corpus (*web for corpus*), es aquella en la que la red se utiliza como fuente de textos en formato electrónico para la

posterior compilación de corpus y emplearlos sin estar en línea (López Rodríguez y Buendía Castro, 2011).

> Por su parte, la web como corpus (*web as corpus*), que utiliza la red directamente como si fuera un corpus propiamente dicho (Baroni y Bernardini, 2006).

El primer enfoque, el de web para corpus, constituye la metodología tradicional de compilación de corpus virtuales que se ha empleado en CULT desde sus inicios. En este enfoque prima la calidad y se busca que los corpus compilados cumplan los requisitos de tamaño, selección, autenticidad y representatividad[2] especificados por McEnery y Wilson (2001).

En lo que atañe al segundo enfoque, el de la web como corpus, la primera pregunta que surge es la siguiente: ¿cumple la red los requisitos que ha de tener una colección de textos para poder considerarse un corpus? Al respecto, Gatto (2013) considera que, a pesar de que Internet no es un corpus en sentido estricto, sí que posee un enorme potencial si se explota desde la perspectiva de la Lingüística de corpus; además, esta autora que constituye una fuente inagotable de lenguaje auténtico que merece estudiar.

Siguiendo la clasificación propuesta por Bernardini et al. (2006), resumida por López Rodríguez y Buendía Castro (2011), consideramos que existen tres maneras de acercarse a la red como corpus desde una perspectiva lingüística:

1. Web como sustituta del corpus (*web as corpus surrogate*). En esta perspectiva se considera la web en sí misma como un gran corpus y se hace uso de la misma con sistemas informáticos que ofrecen una interfaz de búsqueda para posteriormente devolver los resultados a modo de concordancias. Algunas de las herramientas de concordancia en línea más conocidos son WebCorp[3], WebCorp Linguist's Search Engine[4], TAporWare[5] o Corpeus[6].
2. Red como megacorpus o miniweb. Como señalan López Rodríguez y Buendía Castro (2011, pág. 5), se trata de un "intento de crear un nuevo objetivo, una especie de miniweb o megacorpus adaptado a la investigación lingüística". Para explicarlo mejor, podríamos decir que se trataría de crear o concebir un motor

2 Para profundizar en el concepto de representatividad en corpus, remitimos a Seghiri (2011 y 2015).

3 La URL de WebCorp es <http://www.webcorp.org.uk/live/>.

4 Es la versión de WebCorp creada para lingüistas y se puede consultar en <http://wse1.webcorp.org.uk/>.

5 Disponible en <http://taporware.ualberta.ca/>.

6 Se trata de una herramienta de concordancia para el euskera. Se puede acceder al mismo desde <http://corpeus.elhuyar.eus/cgi-bin/kontsulta.py>.

de búsqueda para lingüistas (Fletcher, 2004) con un doble propósito: permitir investigar aspectos lingüísticos en Internet y, al mismo tiempo, permitir investigar aspectos de la Red a través de las lenguas (Bernardini et al., 2006). Un ejemplo de este tipo de motor de búsqueda es el proyecto WaCky[7] (Baroni et al., 2009), desarrollado en la Universidad de Bolonia-Forlì, y que contiene corpus de grandes dimensiones compilados a partir de textos electrónicos para las lenguas inglesa, francesa, alemana e italiana.

3. Web como supermercado de corpus. Es la perspectiva que siguen principalmente los traductores y que concibe la red como si fuera un supermercado en el que se pueden adquirir diferentes corpus. Guarda muchas similitudes con la metodología de compilación de corpus tradicional o web para corpus, aunque existe una diferencia importante entre ambas metodologías: si en la metodología tradicional las distintas fases del protocolo de compilación, a saber, búsqueda y acceso de la documentación, descarga de datos, formato y almacenamiento (Seghiri, 2011 y 2017), se realizan exclusivamente de forma manual, en la metodología que considera la web como supermercado de corpus las fases iniciales del protocolo de compilación (principalmente búsqueda de información y descarga de datos) se realizan de forma semiautomática gracias a las distintas herramientas disponibles. Como hemos especificado en la introducción, en nuestro trabajo optaremos por esta perspectiva al considerar que puede aportar numerosas ventajas a la labor traductológica.

2.2. La compilación automática y semiautomática de corpus

Gracias a la gran ayuda que suponen los corpus tanto paralelos como comparables en el Procesamiento del Lenguaje Natural y en Lingüística Computacional, se han desarrollado herramientas de diverso tipo (Costa et al., 2015) con el fin de crear corpus de forma automática y aprovechar así la fuente inagotable que es la red.

En lo que atañe a corpus paralelos, el objetivo principal es recabar de forma automática *bitextos* o textos paralelos (textos originales y sus correspondientes traducciones a una o varias lenguas). La herramienta pionera creada con dicho propósito fue STRAND[8] (*Structural Translation Recognition Acquiring Natural Data*), a la cual sucedieron otras como WeBiText (Désilets et al., 2008), Bitextor[9] (Esplà-Gomis y Forcada, 2010) e ILSP-FC (Papavassiliou et al., 2013).

7 La URL del proyecto Wacky es <http://wacky.sslmit.unibo.it/doku.php>.

8 Se puede acceder a la misma desde <http://www.umiacs.umd.edu/~resnik/strand/>.

9 Disponible en <https://sourceforge.net/p/bitextor/wiki/Home/>.

Respecto de los corpus comparables, cabe apuntar que la mayoría de las herramientas existentes ofrecen compilación semiautomática en la que existe intervención humana en el proceso de compilación, intervención que suele implicar la filtración de dominios y la eliminación de páginas web (Gutiérrez Florido et al., 2013). De esta forma, la compilación semiautomática supone una clara ventaja con respecto a la compilación automática ya que se evita el ruido documental, esto es, la presencia de información irrelevante o de escasa calidad en el corpus (Costa et al., 2016), de ahí que hayamos optado por esta opción en nuestro trabajo. A pesar de que se han desarrollado herramientas relevantes como Terminus[10], Corpógrafo[11], Babouk (de Groc, 2011) e iCorpora (Costa et al., 2014), resulta indudable que las dos herramientas más conocidas y de uso más extendido son BootCat[12] (Baroni y Bernardini, 2004) y WebBootCat (Baroni et al., 2006). La diferencia entre ambas es que WebBootCat se usa en línea a través de Sketch Engine y BootCat se emplea mediante la instalación de un programa; sin embargo, ambas son prácticamente idénticas en su funcionamiento, ya que se basan en la herramienta Corpus Architect[13]. Dado que en nuestro estudio emplearemos WebBootCat, nos centraremos en explicar esta herramienta en el siguiente apartado.

3. Compilación semiautomática del corpus GeneCorp con WebBootCat

Antes de proceder a detallar cómo se ha compilado el corpus *GeneCorp*, resulta preciso especificar el texto objeto de traducción a fin de comprender por qué se ha compilado el corpus con unas determinadas características.

El encargo presentado en el aula es la traducción del francés al español del artículo "CRISPR-Cas9, l'outil qui révolutionne la génétique"[14], publicado en el número 456 de la revista divulgativa *Pour la Science* con fecha de octubre de 2015. Contiene siete páginas que combinan imágenes y figuras con texto. Pese a ser un artículo divulgativo, cabe señalar que el grado de especialización del artículo es alto ya que contiene numerosa terminología y fraseología sobre genética

10 Accesible desde <http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl>.

11 La URL de Corpógrafo es <https://www.linguateca.pt/corpografo/>.

12 Es el acrónimo de *bootstrapping corpora and terms from the web* y se encuentra disponible en <http://bootcat.dipintra.it>/.

13 Consúltese la URL <https://www.sketchengine.co.uk/my_keywords/corpus-architect/> para conocer más detalles acerca de Corpus Architect.

14 Disponible parte del artículo en la URL <https://www.pourlascience.fr/sd/genetique/crispr-cas-9-loutil-qui-revolutionne-la-genetique-8680.php>.

muy especializada (*doigts de zinc*, *nucléase à motifs TALE, plantes polyploïdes*, *myopathie de Duchenne*, *tyrosinémie*, *plasmide*, etc.). En cuanto a la temática, la modificación genética gracias a la herramienta molecular CRISPR-Cas9 se trata de un tema sobre el que no resulta complicado encontrar documentación al haberse difundido bastante en foros científicos tanto especializados como divulgativos.

Con esta idea en mente, el objetivo es la compilación del corpus *GeneCorp*, el cual se compone a su vez de dos subcorpus: *GeneCorpES*, el subcorpus en español, y *GeneCorpFR*, el subcorpus en francés. En cuanto a los parámetros de diseño, se trata de un corpus virtual o *ad hoc,* bilingüe (francés-español) y comparable. Optamos por compilar un corpus comparable porque coincidimos con Sanz Vicente (2008) en que los corpus comparables representan de forma más real y fiel el comportamiento y uso de cada lengua, ya que, por el contrario, el comportamiento de la lengua de los textos traducidos de los corpus paralelos se puede ver influido. Cabe apuntar también que, en lo que atañe a los límites diasistemáticos (Seghiri, 2011) del corpus *GeneCorp*, únicamente resulta de interés determinar los límites diatópicos, que han sido, en concreto, el español de la variante peninsular para el subcorpus en español y el francés de Francia para el subcorpus en francés.

Respecto de la compilación del corpus *GeneCorp*, procedemos a explicar cómo se ha llevado a cabo de forma semiautomática con WebBootCat. No obstante, explicaremos antes, brevemente, cómo funciona este programa.

Cuando se procede a compilar el corpus de forma semiautomática con WebBootCat, el traductor tiene la posibilidad de acotar el tema del corpus mediante tres opciones: 1) facilitando términos o lemas que definan el tema y que se denominan *seed words*, términos que también pueden obtenerse con la ayuda de Wikipedia, por ejemplo; 2) proporcionando una lista de direcciones URL desde la cual el programa se descarga de forma automática el corpus; y 3) mediante la descarga del contenido completo de una página web. Cabe apuntar que resulta imprescindible especificar la lengua del corpus, ya que WebBootCat únicamente funciona con lenguas que se puedan segmentar de forma automática. Sin embargo, esto no supone un problema dado que Sketch Engine puede trabajar con más de 85 lenguas, entre las cuales están contemplados tanto el francés como el español. WebBootCat ofrece, asimismo, opciones avanzadas de compilación del corpus que se refieren a: 1) realizar la búsqueda a través del buscador Bing; 2) restringir el tamaño del corpus al especificar el número mínimo y máximo de archivos; 3) añadir dos listados: uno de los términos que han de aparecer en el corpus (*white list keywords*) y otro de los términos que no deben aparecer en el corpus (*black list keywords*), que suele corresponderse con las palabras contenidas en una *stop list* o lista de palabras vacías (artículos, adverbios, preposiciones,

etc.)[15]. Una vez compilado el corpus, el traductor puede añadir archivos que se haya descargado previamente mediante la compilación tradicional, lo cual le permite ampliar el corpus y agregar documentos que sean de relevancia.

En el caso del corpus *GeneCorp*, tanto para el subcorpus español como para el subcorpus francés, se combinaron principalmente dos de las opciones de WeBootCat: la opción de facilitar *seed words* y la opción de proporcionar URL. En primer lugar, se empleó la opción de proporcionar *seed words* en ambas lenguas (*CRISPR-Cas9*, *édition génétique*, *ciseaux moléculaires*, *bactéries*, *bactériophage*s, *agressions virales*, *enzyme*, *ARN guide* para el subcorpus *GeneCorpFR* y sus correspondientes equivalentes para el subcorpus *GeneCorpES* (CRISPR-Cas9, edición genética, tijeras moleculares, bacterias, bacteriófagos, agresiones virales, enzima, ARN guía); a su vez, se seleccionaron aquellas URL proporcionadas por WebBootCat que procedían de fuentes fiables y de calidad, como, por ejemplo, un artículo de la revista francesa de divulgación *CNRS Le Journal* titulado "CRISPR-Cas9: des ciseaux génétiques pour le cerveau"[16] o un artículo de la revista española *Genética Médica* llamado "CRISPR para modificar la actividad de los genes sin cambiar el ADN"[17]. Una vez que se habían compilado los dos subcorpus, consideramos oportuno ampliar ambos añadiendo diversas URL obtenidas empleando estrategias de búsquedas concretas como operadores básicos, expresiones booleanas y operadores filtro (Sánchez Ramos, 2017b), siendo de especial relevancia en nuestro caso restringir el dominio a "site:.es" o "site:.fr" para la delimitación diatópica. Respecto de los tipos de textos seleccionados, combinamos tanto artículos divulgativos con alto nivel de especialización como artículos especializados en la materia.

Dado que la compilación es semiautomática, en nuestro caso no fue necesario continuar con las fases de descarga de datos, formato y almacenamiento (Seghiri, 2011 y 2017), ya que WebBootCat se encarga de hacerlo. No obstante, sí que resulta necesario configurar el corpus. Para ello, una vez que ya tenemos seleccionados los textos que formarán parte del corpus, WebBootCat indica que se deben seleccionar dos elementos de análisis: por un lado, el analizador léxico o tokenizador, que distinguirá cada componente léxico (*token*) del texto

15 Tal y como indica Sánchez Ramos (2017a), existen URL desde las que se pueden descargar *stop lists* como <http://www.lextek.com/manuals/onix/stopwords1.html> o <https://www.ranks.nl/stopwords>.

16 El artículo puede consultarse en la siguiente URL: <https://lejournal.cnrs.fr/articles/crispr-cas9-des-ciseaux-genetiques-pour-le-cerveau>.

17 La URL para consultar el artículo es <https://revistageneticamedica.com/2018/01/08/crispr-epigenoma/>.

del corpus y, por otro lado, el analizador gramatical (*TreeTagger*) que lleva a cabo una detección y un estudio automatizado de categorías gramaticales. Una vez que hemos seleccionado estas opciones que nos proporciona WebBootCat, seguidamente la herramienta compila nuestro corpus y nos indica que ya está listo para la fase de explotación.

Así, tras la fase de compilación semiautomática, el corpus *GeneCorp* se compone de los dos subcorpus siguientes: *GeneCorpFR*, el subcorpus francés, con 50 800 palabras (o *tokens*) y *GeneCorpES*, el subcorpus español, compuesto por 49 706 palabras (o *tokens*). Somos conscientes que se trata de un corpus de reducido tamaño, pero hemos preferido cribar los textos que lo integran en aras de la calidad.

4. Explotación del corpus GeneCorp con la herramienta de análisis de corpus Sketch Engine

Antes de proceder a explicar cómo se ha explotado el corpus *GeneCorp* en Sketch Engine, consideramos oportuno acercarnos a este programa y describirlo brevemente. Sketch Engine constituye una herramienta de análisis y gestión de corpus creada por Kilgariff (Kilgarriff et al., 2014) cuyo uso está cada vez más extendido entre lingüistas. El sistema se basa en el sistema de gestión de corpus Manatee (Rychlý, 2007), el cual permite gestionar con eficacia corpus de gran tamaño. Asimismo, emplea la interfaz gráfica de usuario Bonito (Rychlý, 2007).

A continuación, especificaremos las funcionalidades de Sketch Engine que nos han resultado más interesantes a la hora de explotar el corpus *GeneCorp*. Para este apartado hemos seguido en parte la metodología que propone Sánchez Ramos (2017a), si bien hemos organizado los beneficios que supone el uso de Sketch Engine dividiendo el apartado en las funcionalidades más relevantes para los traductores a la hora de explotar un corpus: a) Sketch Engine como herramienta terminológica; b) Sketch Engine como programa de concordancias; c) Sketch Engine para el análisis de comportamientos gramaticales y colocacionales; y d) Sketch Engine para la creación de tesauros distribucionales.

4.1. Sketch Engine como herramienta terminológica

Sketch Engine permite generar listas de términos ordenados por orden alfabético y por frecuencia de aparición en su función *Word lists*. Para que esta opción arroje resultados pertinentes, resulta imprescindible activar la opción de *black list keywords* con objeto de evitar que aparezcan palabras vacías en las listas. Presentamos en la Figura 1 las listas monoléxicas del subcorpus *GeneCorpFR* y del subcorpus *GeneCorpES,* respectivamente.

Figura 1: Listas de unidades terminológicas monoléxicas más frecuentes en GeneCorpFR (izquierda) *y* GeneCorpES (derecha)

word	frequency
ADN	122
Cas	74
CRISPR	73
CRISPR-Cas	72
génétique	62
ARN	57
cellules	55
séquence	48
gène	46
génome	44
gènes	42
enzyme	36
permet	33
système	31
technique	30
outil	29
virus	28
souris	26
bactéries	25
génétiques	25
modifier	25
séquences	24
chercheurs	24
utilisation	20
cellule	20
protéine	20
édition	20
modification	19
équipe	18
faire	17
guide	17
université	16
années	16
recherche	16
Leishmania	15
mutations	15
maladies	15
déjà	15
exemple	15
homme	15

word	frequency
células	192
CRISPR	152
ADN	152
genoma	112
sistema	105
secuencia	103
gen	100
forma	88
ARN	86
genética	79
genes	76
edición	76
bacterias	73
virus	71
técnica	71
CRISPR-Cas9	65
investigación	58
derecho	57
modificación	55
secuencias	54
tecnología	52
enfermedades	50
humanos	50
célula	47
Cas9	46
proteína	46
génica	45
enzima	44
proteínas	44
derechos	43
herramienta	43
CRISPR/Cas9	42
embriones	41
modificar	41
CSK	40
genético	39
uso	38
mutaciones	35
tipo	35
vida	35

Comparando ambas listas ya encontramos equivalentes en español para términos franceses como, por ejemplo, *cellules*/células, *séquence*/secuencia, *ARN*/ARN, *gène*/gen, *bactéries*/bacterias, *mutation*/mutación, *enzyme*/enzima, *CRISPR-Cas9*/CRISPR/Cas9 o *protéine*/proteína.

Además de la lista de unidades terminológicas monoléxicas, Sketch Engine ofrece también la posibilidad de ampliar el número de gramas, creando así listas de unidades terminológicas poliléxicas. Se trata de una característica que, como apunta Sánchez Ramos (2017a), resulta de gran utilidad para la creación de glosarios. Mostramos en la Figura 2 las listas de unidades terminológicas poliléxicas del subcorpus *GeneCorpFR* y del subcorpus *GeneCorpES,* respectivamente.

Figura 2: Listas de unidades terminológicas poliléxicas de GeneCorpFR (izquierda) *y* GeneCorpES (derecha)

word (n-grams)	frequency
enzyme Cas	19
cellules souches	13
ARN guide	11
souches embryonnaires	9
cellules souches embryonnaires	9
système CRISPR	8
prix Nobel	8
forçage génétique	8
ciseaux génétiques	8
ARN guides	8
édition génomique	7
édition génétique	6
technologie CRISPR	6
recombinaison homologue	6
petit ARN	6
ADN viral	6
séquences CRISPR	5
souches embryonnaires humaines	5
recherche fondamentale	5
maladie génétique	5
génie génétique	5
embryonnaires humaines	5
complexe CRISPR-Cas	5
cellules souches embryonnaires humaines	5
DNA replication	5
ADN étranger	5
thérapie génique	4
séquence homologue	4
système CRISPR-Cas	4
outil génétique	4
nouvel outil	4
modifications génétiques	4
modification génétique	4
maladies génétiques	4
devient possible	4
cellules humaines	4
êtres vivants	3
éthiques soulevés	3
utilisant CRISPR-Cas	3
technique révolutionnaire	3
souris portant	3
souches embryonnaires humaines exprimant	3
replication dynamics	3
questions éthiques	3
protéine Cas	3
nouvel ADN	3
matériel génétique	3
marqueur fluorescent	3
long ARN	3
humaines exprimant	3
génétiques CRISPR-Cas	3
génétiquement modifiés	3

word (n-grams)	frequency
recombinación homóloga	27
edición genómica	25
modificación génica	23
material genético	18
ARN guía	18
sistema CRISPR	17
derechos fundamentales	17
sistema CRISPR-Cas9	16
enzima Cas9	16
edición genética	16
embriones humanos	15
células humanas	15
investigación científica	14
vector donante	13
sistema CRISPR/Cas9	13
progreso científico	13
ingeniería genética	13
genome editing	11
gen csk	11
Oliver Smithies	11
patrimonio genético	10
modificación genética	10
línea germinal	10
génica dirigida	10
gene targeting	10
derecho fundamental	10
terapia génica	9
sistema CRISPR-Cas	9
medio ambiente	9
línea celular	9
distrofia muscular	9
dignidad humana	9
células Jurkat	9
Instituto Broad	9
técnica CRISPR	8
tijeras moleculares	8
sistema inmune	8
secuencias CRISPR	8
propiedad intelectual	8
premio Nobel	8
modificación génica dirigida	8
derechos humanos	8
células heterocigotas	8
ADN extraño	8
tipos celulares	7
sistemas CRISPR	7
embryonic stem	7
proteínas Cas	6
modelos animales	6
gene editing	6
comunidad científica	6
Sánchez Amat	6

Al igual que con las unidades monoléxicas, gracias a estas listas ya es posible encontrar equivalentes en español a unidades terminológicas poliléxicas en francés, como son las siguientes: *enzyme Cas*/enzima Cas9, *ARN guide*/ARN guía, *système CRISPR*/ sistema CRISPR, *modification génétique*/modificación genética, *édition génétique*/edición genética, *recombinaison homologue*/recombinación homóloga, *ciseaux moléculaires*/tijeras moleculares, *forçage génétique*/genética dirigida o *ADN étranger*/ADN extraño.

Otra opción terminológica interesante que ofrece Sketch Engine es extraer mediante la función *Keywords/terms* los términos clave. Para ello, se tiene en cuenta el *keyness score*, es decir, una puntuación que indica la frecuencia de un término en el corpus objeto de estudio con respecto al corpus de referencia. Para *GeneCorpFR* el corpus de referencia es French Web 2012 y para *GeneCorpES*, el corpus Spanish Web 2011. En las Figuras 3 y 4 se muestran las listas resultantes para nuestros dos subcorpus, listas en las que consta la siguiente información: *score* es el *keyness score* del término en cuestión; *F* es la frecuencia en el corpus objeto de estudio (*GeneCorpFR* y *GeneCorpES*) y *RefF* indica la frecuencia en el corpus de referencia.

Figura 3: Listas de términos clave de GeneCorpFR *extraídos mediante* keyness score

Single-word	Score	F	RefF
CRISPR	4,377.09	73	21
Cas	1,910.81	70	13,741
ARN	1,214.48	57	20,824
nucléase	943.60	16	222
génome	861.68	51	29,250
Leishmania	763.59	13	272
ADN	681.75	119	108,549
enzyme	665.87	36	25,733
Doudna	601.51	10	1
replication	571.41	10	604
Charpentier	471.55	14	8,986
Causeret	420.11	7	35
recombinaison	404.83	9	3,864
Orc	393.00	9	4,325
tracr	361.24	6	3
inactiver	355.20	9	6,003
forçage	354.51	8	4,098
bactériophage	333.59	6	952
génomique	326.79	10	9,623
généticien	316.10	9	8,161
gène	312.99	88	181,845
réplication	312.93	12	14,949
palindrome	308.81	6	1,947
Chneiweiss	298.70	5	99
genome	285.02	5	653
génétique	284.09	90	206,348
génique	273.09	8	8,732
DNA	267.57	9	11,717
crible	259.98	10	15,037
nucléotide	256.20	6	4,697
embryonnaire	244.70	12	22,309
TALEN	241.12	4	5
pluripotence	237.78	4	166
Morizot	234.64	4	321
Cdc	231.34	4	489
ciseau	228.88	18	42,660
Inserm	228.15	8	12,707

Multi-word	Score	F	RefF
doigts de zinc	596.79	10	96
ciseaux génétiques	481.45	8	0
forçage génétique	421.39	7	0
édition du génome	421.39	7	0
édition génomique	421.39	7	0
cellules souches embryonnaires	392.11	9	0
souches embryonnaires	390.70	9	4,424
édition génétique	359.18	6	0
modifications génétiques	327.60	6	1,190
complexe crispr-cas	301.28	5	0
cellules souches embryonnaires humaines	281.31	5	0
souches embryonnaires humaines	281.31	5	824
outil génétique	240.50	4	0
analyse bioinformatique	240.26	4	0
this tool	240.26	4	0
cellules souches	231.49	13	0
génie génétique	216.28	5	4,507
modification génétique	215.38	4	1,379
maladie génétique	213.37	5	4,719
cellules humaines	200.69	4	0
in vitro	185.44	10	0
replication dynamics	181.17	3	0
marqueur fluorescent de la pluripotence	181.17	3	0
ciseaux moléculaires	180.63	3	0
marqueur fluorescent	179.91	3	90
recombinaison homologue	177.79	3	0
cellules de mammifères	177.27	3	0
ensemble des gènes	177.10	3	0
technique révolutionnaire	175.72	3	0
attaque virale	175.55	3	0
maladies génétiques	175.31	4	4,309
cellules du foie	169.95	3	0
thérapie génique	168.22	4	4,977
us to	167.91	3	0
embryons humains	155.78	3	0
embryon humain	154.98	3	0
recherche fondamentale	143.54	5	0

Mostramos, a continuación, la Figura 4.

Figura 4: Listas de términos clave de GeneCorpES *extraídos mediante* keyness score

Single-word	Score	F	RefF
crispr	3,098.94	154	1
csk	1,164.66	59	220
cas9	925.68	46	9
arn	719.65	86	15,441
nucleasas	613.28	31	204
blasticidina	564.26	28	1
genoma	520.84	125	42,077
génico	512.11	55	12,772
recombinación	499.14	34	4,091
genome	484.72	26	892
heterocigotas	468.16	24	368
charpentier	446.64	24	915
montoliu	407.43	21	433
lyp	401.87	20	41
primers	400.48	22	1,183
editing	390.46	20	363
doudna	382.97	19	8
dna	375.14	53	20,268
plásmidos	352.67	21	2,206
jurkat	341.33	17	54
ligación	333.39	17	317
gene	311.68	34	13,159
plásmido	300.45	17	1,556
zhang	298.34	25	7,571
capecchi	295.43	15	273
rna	269.96	22	7,066
µl	264.98	18	4,069
targeting	258.88	14	1,009
knockout	254.99	14	1,192
p1	251.01	18	4,907
genómica	250.79	26	11,971
cell	248.75	25	11,271
oligonucleótidos	238.82	13	1,091
talar	229.30	48	35,328
alelo	226.81	23	11,475
nucleofección	222.30	11	0
adn	222.21	152	140,241

Multi-word	Score	F	RefF
modificación génica	543.65	27	3
recombinación homóloga	529.89	27	64
edición genómica	503.96	25	0
sistema crispr	483.84	24	0
arn guía	383.25	19	0
célula heterocigotas	343.01	17	0
enzima cas9	322.89	16	0
sistema crispr-cas9	322.89	16	0
edición genética	321.93	16	8
wild type	302.47	15	3
revista iberoamericana de bioética	282.66	14	0
sistema crispr-cas	282.66	14	0
vector donante	282.66	14	0
mario capecchi	261.49	13	10
célula humana	239.92	15	615
técnica de edición	234.22	12	83
embrión humano	231.29	17	1,133
cassette de blasticidina	222.30	11	0
emmanuelle charpentier	222.30	11	0
secuencia crispr	222.30	11	0
dedo de zinc	220.76	11	18
revista iberoamericana	220.48	14	662
oliver smithies	220.10	11	25
línea germinal	190.02	10	151
material genético	187.47	18	2,198
célula jurkat	182.06	9	2
patrimonio genético	181.17	10	272
martin evans	181.16	9	14
instituto broad	180.08	9	27
distrofia muscular	178.78	12	836
enzima de restricción	175.06	9	94
progreso científico	174.21	13	1,189
modificación genética	173.40	11	663
nobel de fisiología	169.36	9	177
tipo celular	166.64	11	783
proteína ca	161.95	8	0
técnica crispr	161.95	8	0

A pesar de que los listados de palabras clave nos permiten también obtener equivalentes de los términos monoléxicos y poliléxicos del texto origen (*nucléase*/nucleasa, *doigts de zinc*/dedos de zinc, *édition génomique*/edición genómica), cabe apuntar que Sketch Engine no permite el empleo de listas de palabras vacías en la función *Keywords/terms* y, en consecuencia, encontramos más ruido que en los listados obtenidos a través de la opción de listas de palabras (*Word list*).

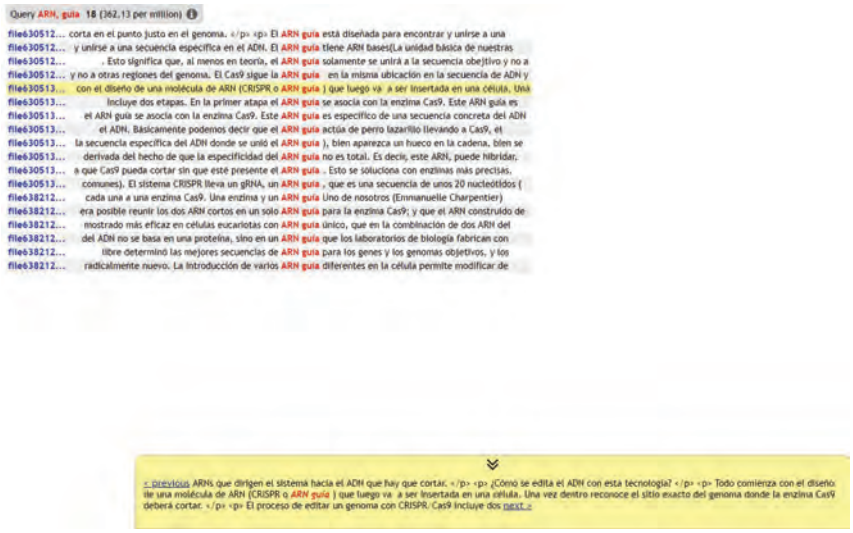
4.2. Sketch Engine como programa de concordancias

Las listas de términos obtenidas tanto a través de la opción de lista de palabras (*Word list*) como de la opción de términos clave (*Keywords/terms*) permiten explorar el comportamiento de las unidades terminológicas monoléxicas o poliléxicas en contexto gracias a la función de concordancia. Así, en la Figura 5 mostramos las concordancias para *ARN guide* en el subcorpus *GeneCorpFR*.

Figura 5: Líneas de concordancias para ARN guide *en el subcorpus* GeneCorpFR

Por su parte, en la Figura 6, se recogen las concordancias de "ARN guía" en el subcorpus *GeneCorpES*:

Figura 6: Líneas de concordancias para ARN guía *en el subcorpus* GeneCorpES



Otra utilidad que permite la generación de concordancias es consultar información conceptual sobre términos o palabras clave. Para ello es necesario introducir patrones de búsqueda, normalmente con comodines. En nuestro caso, dado que en el texto origen aparecen distintos tipos de células, hemos querido comprobar qué información arroja cada subcorpus al respecto, así que, dentro de la opción de búsqueda (*Search*), empleamos los siguientes patrones de búsqueda: "cellule *" para *GeneCorpFR* y "célula *" en *GeneCorpES*. Los resultados muestran que existen muchos más tipos de células en *GeneCorpES*, en concreto 25 tipos ("célula inmune", "células adultas", "células embrionarias", "células germinales", "células vivas", "células sexuales", "células humanas", "células somáticas", "células bacterianas", "células embrionarias", "células madre", "células hepáticas", "células pancreáticas", "células somáticas", "células eucariotas", "células huésped", "células epiteliales", "célula diana", "célula hospedadora", "célula procariota", "célula Jurkat", "células T", "células heterocigotas", "células homocigotas" y "células troncales"); sin embargo, en *GeneCorpFR* encontramos únicamente 11 tipos (*cellules cibles*, *cellules sexuelles*, *cellules humaines*, *cellules germinales*, *cellules vivantes*, *cellule animale*, *cellules embryonnaires*, *cellules souches*, *cellules œufs*, *cellule séquence* y *cellules immunitaires*). No obstante, a pesar de la diferencia en el número de tipos de células, es posible establecer paralelismos entre tipos de células equivalentes en francés y español como son *cellule cible* y "célula diana", *cellules germinales* y células germinales, *cellules embryonnaires* y "células embrionarias", *cellules souches* y "células madre", *cellules sexuelles* y "células sexuales", *cellules humaines* y "células humanas" y, por último, *cellules vivantes* y "células vivas".

La búsqueda de concordancias ofrece la posibilidad de analizar colocaciones y patrones colocacionales o fraseológicos a través de la función *Collocations*. Sketch Engine permite seleccionar, por un lado, el rango de palabras del patrón colocacional (es decir, las palabras que aparecen a la izquierda y a la derecha de la palabra clave) y, por otro, la frecuencia mínima de aparición de la colocación en cuestión. En nuestro caso, consideramos de interés conocer el patrón del lexema *cibl** (lo cual incluye, entre otros, el *verbo cibler*, los participios *ciblé*, *ciblée*, *ciblés* y *ciblées* y el sustantivo *ciblage*) ya que se trata de una forma verbal de frecuente aparición en el texto origen. En la Figura 7 mostramos parte del patrón colocacional de *cibl**.

Figura 7: Patrón colocacional de cibl* *en el subcorpus* GeneCorpFR

Collocation candidates

		Cooccurrence count	Candidate count	T-score	MI	logDice
P N	ADN	11	122	3.241	5.464	11.174
P N	génome	5	44	2.195	5.798	11.036
P N	gène	5	46	2.194	5.734	11.000
P N	gènes	4	42	1.957	5.543	10.752
P N	d'	6	104	2.362	4.820	10.476
P N	séquence	3	48	1.675	4.935	10.227
P N	et	9	227	2.845	4.279	10.142
P N	cellules	3	55	1.667	4.739	10.109
P N	que	4	91	1.907	4.428	10.034
P N	l'	6	158	2.317	4.217	10.000
P N	des	10	300	2.968	4.028	9.938
P N	dans	5	133	2.114	4.202	9.938
P N	l'	6	174	2.304	4.077	9.884
P N	un	6	190	2.291	3.950	9.777
P N	du	4	117	1.880	4.065	9.761

El análisis del patrón colocacional permite, además, comprobar qué lemas aparecen junto al término objeto de búsqueda mediante dos filtros, tal y como puede comprobarse en la Figura 8.

Figura 8: Patrón colocacional de cibl* *con filtro positivo y negativo con* ADN *en* GeneCorpFR

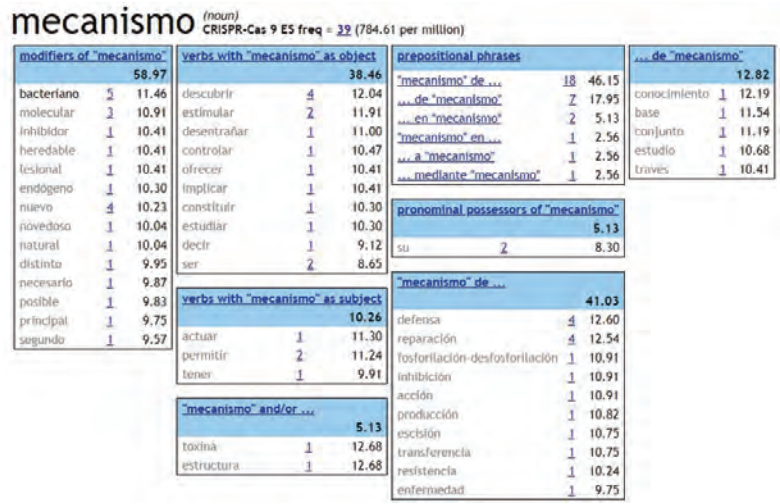
Como puede comprobarse, la opción *P* (*positive filter*) o filtro positivo se comprueba el término en contexto incluyendo también el lema que lo acompaña, que en nuestro caso es *ADN* (parte superior de la Figura 8), mientras que la opción *N* (*negative filter*) o filtro negativo permite comprobar el término en contexto, pero sin el lema que lo acompaña, es decir, excluyendo, en nuestro caso, *ADN* (parte inferior de la Figura 8).

4.3. Sketch Engine para el análisis de comportamientos gramaticales y colocacionales

Sketch Engine ofrece también la posibilidad de explorar los patrones gramaticales, colocacionales y sintácticos de lemas gracias a dos funciones específicas: por un lado, la función *Word sketch* o comportamiento gramatical, colocacional y sintáctico de un único término o lema y, por otro, la función *Sketch diff*, que permite comparar el comportamiento gramatical, colocacional y sintáctico de dos términos o lemas según ocurren en el corpus.

La función *Word sketch* permite explorar los patrones sintácticos de un lema o término concreto proporcionando información sobre la posición y función sintáctica. En nuestro caso, nos centraremos en el lema "mecanismo", que aparece en 39 ocasiones en el subcorpus *GeneCorpES* y cuyo patrón sintáctico se puede comprobar en la Figura 9.

Figura 9: Word sketch *de* mecanismo



Además, *Word sketch* tiene otra función que le puede ser de gran utilidad al traductor: existe la posibilidad de presentar un patrón sintáctico bilingüe en la que aparezcan los dos equivalentes. En la Figura 10, mostramos el patrón sintáctico de *mécanisme* y de "mecanismo".

Figura 10: Word sketch *bilingüe de* mecanismo *y* mécanisme

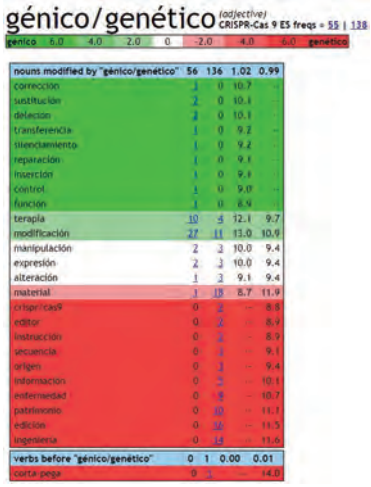
En lo que respecta a *Sketch diff*, su interés principal radica en que, como ya apuntamos, es posible comparar los patrones sintácticos, gramaticales y colocacionales de dos lemas o términos, función que puede ser de gran ayuda a la hora

de traducir. En la traducción que nos ocupa, se nos ha planteado la duda de si, en determinados casos, se ha de emplear "génico" o "genético" en la traducción al español. Por este motivo, decidimos emplear *Sketch diff* con objeto de comparar el comportamiento de ambos adjetivos en los dos subcorpus.

Como se indica en la Figura 11, en el subcorpus *GeneCorpES* "génico" aparece 55 veces, mientras que "genético" se recoge en 138 ocasiones. Los colores verde y rojo se corresponden con cada uno de los adjetivos introducidos, a saber, verde con "génico" y rojo con "genético". En consecuencia, comprobamos que los sustantivos "corrección", "sustitución", "deleción", "transferencia", "silenciamiento", "reparación" "inserción", "control", "función" aparecen con "génico", mientras que los sustantivos "CRISPR-Cas9", "editor", "instrucción", "secuencia", "origen", "información", "enfermedad", "patrimonio", "edición", "ingeniería" coocurren con "genético". El grado de degradación del color se asocia a la probabilidad de compartir patrones. Cuando se degrada el color significa que la colocación es menos cercana o menos típica del adjetivo en cuestión. Así, "terapia" y "modificación" es más común que aparezcan con "génico", mientras que "material" es más frecuente que aparezca con "genético". Los sustantivos en blanco ("manipulación", "expresión" y "alteración") son comunes a ambos adjetivos.

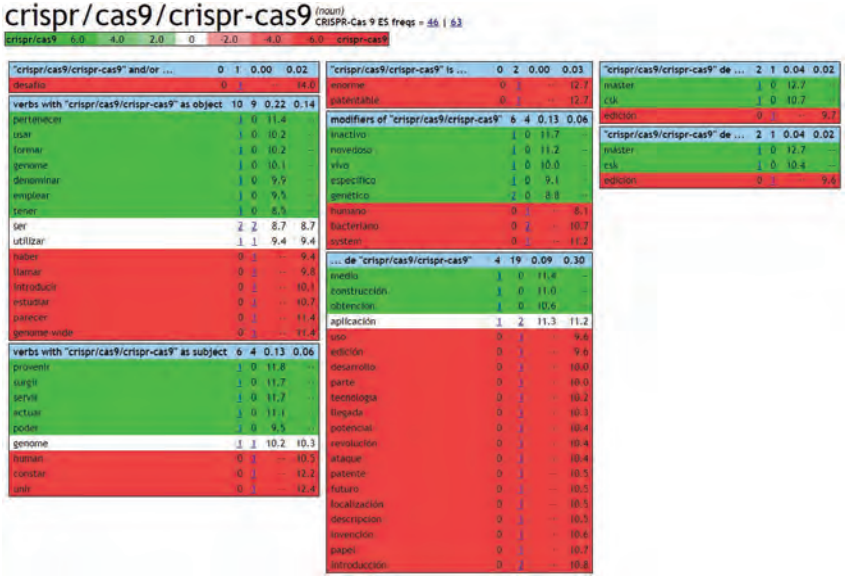
En cuanto a las cifras, las dos primeras (56 y 136) indican la frecuencia de coocurrencia con el primer adjetivo y el segundo, respectivamente. Las últimas dos cifras (1,02 y 0,99) indican el índice de distinción (*salient score*) respecto a cada adjetivo.

Figura 11: Comparación del comportamiento sintáctico y colocacional de génico y genético



Otra comparación del comportamiento sintáctico, gramatical y colocacional que nos ha sido de gran ayuda en la traducción del texto origen es comparar qué grafía es más frecuente en español para CRISPR-Cas9, ya que se presentan dos posibilidades: CRISPR-Cas9 (con guion) y CRISPR/Cas9 (con barra).

Figura 12: Comparación de la grafía de CRISPR-Cas 9 y CRISPR/Cas9 en GeneCorpES



Gracias a *Sketch diff*, hemos podido determinar, como se puede comprobar en la Figura 12, que es más frecuente con guion (aparece en 63 ocasiones frente a las 46 con barra).

4.4. Sketch Engine para la creación de tesauros distribucionales

Sketch Engine ofrece la posibilidad de crear tesauros distribucionales (función *Thesaurus*), definidos por Calvo et al. (2005, pág. 2) como "a thesaurus generated automatically from a corpus by finding words which occur in similar contexts to each other". Se calcula en base a las palabras o lemas que suelen aparecer con las mismas colocaciones que una palabra dada. La Figura 13 contiene el tesauro distribucional para *séquence*, lo que permite comprobar que dicho término tiende a aparecer en contextos similares a *maladie*, *édition*, *modification*, *cellule*, *gène*, *génome*, *outil*, *ciseau*, *équipe* y *mutation*.

Figura 13: Tesauro distribucional para séquence

séquence *(noun)* CRISPR-Cas 9 FR freq = 71 (4,264.00 per million)

Lemma	Score	Freq
équipe	0.152	23
génome	0.151	51
ciseau	0.140	18
modification	0.131	33
mutation	0.129	29
édition	0.125	20
maladie	0.121	29
outil	0.116	34
cellule	0.104	75
gène	0.078	88

Consideramos que, a priori, esta función no resulta de gran utilidad para el traductor, si bien puede ayudarle a ver cómo se relacionan conceptos relacionados, aclarando así posibles dudas conceptuales.

5. Conclusiones

Al inicio de nuestro trabajo nos habíamos propuesto una serie de objetivos que estimamos haber cumplido. En concreto, fueron cuatro objetivos secundarios y uno principal. En cuanto a los objetivos secundarios, el primero de ellos lo hemos cumplido tras explicar la diferencia entre web para corpus y web como corpus, al mismo tiempo que hemos especificado las tres perspectivas posibles dentro de la web como corpus: la web como sustituta del corpus, la red como megacorpus o miniweb y la web como supermercado de corpus. Hemos concretado que, en nuestra propuesta, hemos seguido la perspectiva de la web como supermercado de corpus ya que todas las fases de compilación de nuestro corpus objeto de estudio se han llevado a cabo de forma semiautomática. Corresponde también al primer objetivo el haber realizado un excurso por la compilación automática y semiautomática de corpus, centrándonos en particular en las herramientas que agilizan esta tarea y permiten compilar corpus automática o semiautomáticamente.

Nos centraremos a continuación en el tercer objetivo secundario dado que el segundo objetivo se ha logrado gracias al cumplimiento de los objetivos tercero y cuarto. El tercer objetivo aludía a la compilación del corpus *ad hoc*, comparable y bilingüe (francés-español) sobre genética mediante la función WebBootCat,

disponible a través de Sketch Engine, que ha sido utilizado en la asignatura Traducción científica y técnica de la lengua B (francés) del Grado en Traducción e Interpretación de la Universidad de Córdoba. Tras explicar brevemente el encargo objeto de traducción directa del francés al español, a saber, el artículo "CRISPR-Cas9, l'outil qui révolutionne la génétique", hemos explicado cómo funciona WebBootCat (cumpliendo así en parte el segundo objetivo secundario) y las distintas opciones que ofrece para compilar corpus. Hemos concretado que se ha optado por combinar dos de las opciones de WebBootCat: la opción de emplear *seed words* y la opción de añadir nuestras propias URL tras búsquedas en Internet con estrategias de búsquedas concretas. El resultado final de la compilación es el corpus *GeneCorp*, un corpus compuesto de los dos subcorpus siguientes: *GeneCorpFR*, el subcorpus francés acotado diatópicamente al francés de Francia y con 50 800 palabras *(o tokens) y GeneCorpES*, el subcorpus español, acotado diatópicamente al español de España y compuesto por 49 706 palabras (o *tokens*). Hemos justificado el reducido tamaño del corpus esgrimiendo que primaba la calidad. Consideramos, por lo tanto, que también hemos cumplido el tercer objetivo.

En lo que atañe al cuarto objetivo secundario, consideramos oportuno vincularlo con el objetivo principal de nuestro trabajo, ya que el grueso de este trabajo se ha centrado en presentar cómo puede explotarse un corpus mediante Sketch Engine para facilitar la labor traductológica en el aula de la asignatura Traducción científica y técnica de la lengua B (francés). De esta forma, tras describir brevemente ciertos aspectos técnicos de Sketch Engine (cumpliendo así en parte el segundo objetivo secundario), hemos detallado cómo puede emplearse Sketch Engine para la explotación de corpus, para lo cual hemos seguido en parte a Sánchez Ramos (2017a). No obstante, hemos optado por clarificar los beneficios que aporta la herramienta centrándonos en los aspectos más relevantes para el traductor; para ello, hemos concretado cómo ayuda Sketch Engine 1) en la labor terminológica; 2) en el análisis de concordancias; 3) en el análisis de comportamientos gramaticales y colocacionales; y 4) en la creación de tesauros distribucionales.

En lo que atañe a la labor terminológica, Sketch Engine genera listas de términos ordenados por frecuencia de aparición, ofreciendo la posibilidad de crear tanto listas de unidades terminológicas monoléxicas como de unidades terminológicas poliléxicas. En nuestro caso, gracias al corpus *GeneCorp*, hemos podido obtener los equivalentes en español tanto para términos franceses monoléxicos (*séquence*/secuencia, *ARN*/ARN, *gène*/gen, bactéries/bacterias, etc.) como para términos poliléxicos en francés (*enzyme Cas*/enzima Cas9, *ARN guide*/ARN guía, *système CRISPR*/ sistema CRISPR, *ciseaux molécu*laires/tijeras moleculares, *forçage génétique*/genética dirigida, *ADN étranger*/ADN extraño, entre otros).

Coincidimos con Sánchez Ramos (2017a) en que se trata de una característica que resulta de gran utilidad para la creación de glosarios, motivo por el cual le será de gran utilidad al traductor cuando precise elaborarse su propio material terminológico.

Dentro de la labor terminológica, hemos presentado, a su vez, la opción de extracción de términos mediante *keyness score*. Consideramos que los resultados arrojados por Sketch Engine son de interés para el traductor, ya que le permite también obtener equivalentes de los términos monoléxicos y poliléxicos del texto origen (*nucléase*/nucleasa, *doigts de zinc*/dedos de zinc, *édition génomique*/edición genómica). Sin embargo, al no poder emplear listas de palabras vacías, nos hemos encontrado con más ruido del deseable, por lo que, a pesar de la utilidad de la opción de extracción de términos, nos parece que la opción de lista de palabras resulta más conveniente para el traductor.

En cuanto al análisis de concordancias, hemos explicado que las listas de términos obtenidas tanto a través de la opción de lista de palabras (*Word list*) como de la opción de términos clave permiten explorar el comportamiento de las unidades terminológicas monoléxicas o poliléxicas en contexto gracias a la función de concordancia. Asimismo, hemos explorado la opción de consultar información conceptual sobre palabras clave, centrándonos en particular en los distintos tipos de células que aparecen en nuestro corpus *GeneCorp*. Hemos detectado que son muchos más los tipos de células presentes en el subcorpus *GeneCorpES*, pero, aun así, hemos sido capaces de establecer paralelismo entre tipos de células equivalentes en francés y español como son, entre otras, *cellule cible* y "célula diana", *cellules germinales* y células germinales y *cellules embryonnaires* y "células embrionarias". Para terminar, en el apartado de análisis de concordancias, hemos ahondado en el estudio del patrón colocacional de la forma verbal *cibl** a través de la función *Collocations*.

Respecto del análisis de comportamientos gramaticales y colocacionales, nos hemos centrado en dos de las funciones más características de Sketch Engine: por un lado, la función *Word sketch* o comportamiento gramatical, colocacional y sintáctico de un término o lema y, por otro, la función *Sketch diff*, que permite comparar el comportamiento gramatical, colocacional y sintáctico de dos términos o lemas según ocurren en el corpus. Con la función *Word sketch* hemos conocido el patrón sintáctico del lema "mecanismo", no solo en español, sino también de su equivalente en francés (*mécanisme*) gracias a la opción bilingüe de la función. En lo que concierne a *Sketch diff*, en primer lugar, hemos comparado los comportamientos sintácticos y colocacionales de "génico" o "genético", dos adjetivos ampliamente utilizados en el corpus *GeneCorpES* y, en segundo

lugar, hemos aclarado que en español la grafía más frecuente para el término "CRISPR-Cas9" es con guion en lugar de con barra, aspecto ortotipográfico de suma relevancia para el traductor y que, gracias a Sketch Engine, aclararlo supone ahorrar mucho tiempo y esfuerzo.

En el último aspecto abordado en Sketch Engine, la creación de tesauros distribucionales, hemos creado el tesauro para el término *séquence*. Tal y como hemos explicado, no se trata, a priori, de una función de gran utilidad para el traductor, si bien puede ayudarle a ver cómo se relacionan conceptos relacionados y puede servirle para aclararle posibles dudas conceptuales.

En suma, consideramos que emplear Sketch Engine constituye un beneficio claro para el traductor ya que le va a permitir adquirir conocimiento conceptual, terminológico y fraseológico sobre un tema concreto. La metodología CULT en la Lingüística de Corpus avanza y bebe de nuevas fuentes y nuevos planteamientos, de ahí que también se beneficie de nuevas herramientas que aportan numerosas ventajas. Emplear Sketch Engine y WebBootCat junto con corpus *ad hoc* significa beneficiarse de la fundamental aportación que es el corpus en los Estudios de Traducción, para, además, combinar el corpus con una herramienta que, entre otras ventajas, 1) supone un ahorro de tiempo y esfuerzo; 2) agiliza en gran medida la compilación de corpus manteniendo un alto nivel de calidad; 3) facilita la extracción terminológica y la posterior elaboración de glosarios; 4) permite explorar patrones sintácticos, colocacionales y gramaticales que son de gran ayuda al traductor. Por todo lo anterior, estimamos que la combinación de corpus *ad hoc* y Sketch Engine ayuda a obtener recursos terminológicos, fraseológicos y temáticos de calidad, a coste cero y con suma rapidez, con lo que su uso tanto en el aula de Traducción como en la labor profesional del traductor y del intérprete es altamente recomendable.

6. Agradecimientos

El presente trabajo ha sido realizado en el seno de la red temática TRAJUTEC y de la red docente de excelencia TACTRAD (ref. 719/2018), ambas de la Universidad de Málaga, así como en el marco de los proyectos VIP (Ref. FFI2016-75831-P), TERMITUR (Ref. HUM2754), NOVATIC (Ref. PIE15-145, UMA), INTERPRETA 2.0 (Ref. PIE17-015, UMA), PROFETA (Ref. PIE19-033, UMA), UCOTERM (Ref. 2017-1-1005), POSTrad II (Ref. PIE 109, UVa), POSTrad III (Ref. PIE 102, UVa), INGENIO (J.A.), TRIAJE (Ref. UMA18-FEDERJA-067) y PROFETA (Ref. PIE19-033, UMA).

7. Referencias bibliográficas

Baroni, M., & Bernardini, S. (2004). BootCat: Bootstrapping Corpora and Terms from the Web. En M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva (Eds.), *Proceedings of LREC* (pp. 1313–1316). Lisboa: ELRA.

Baroni, M., & Bernardini, S. (Eds.) (2006). *Wacky! Working on the Web as Corpus*. Bolonia: GEDIT.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCat: Instant Domain-Specific Corpora to Support Human Translators. En A. Lynum & L. Korsnes (Eds.), *Proceedings of EAMT 2006- 11th Annual Conference of the European Association for Machine Translation* (pp. 247–252). Oslo: EAMT.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–231.

Beeby, A., Rodríguez Inés, P., & Sánchez Gijón, P. (Eds.) (2009). *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Ámsterdam/Filadelfia: John Benjamins Publishing.

Bernardini, S., Baroni, M., & Evert, S. (2006). A Wacky Introduction. En M. Baroni & S. Bernardini (Eds.), *WaCky! Working Papers on the Web as Corpus* (pp. 1–32). Bolonia: GEDIT.

Borel, J., & Loock, R. (2017). *Ça "corpus" pas mal du côté des logiciels de TAO : l'heure de la convergence aurait-elle sonné ?* Ponencia presentada en el congreso La gestion des contraintes génériques/textuelles par les traducteurs : annotation, modélisation, extraction de l'information dans les corpus électroniques, Dijon, Francia.

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Londres: Routledge.

Calvo, H., Gelbukh, A., & Kilgarriff, A. (2005). Distributional Thesaurus Versus WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment. En A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13–19, 2005, Proceedings* (pp. 172–183). Berlín/Nueva York: Springer.

Castillo Rodríguez, C. M. (2014). Online Sources for a Corpus Compilation Specialized in Wellness and Beauty Tourism: A Brief Approach for Translators' Documentation. En. J. F. Durán Medina & I. Durán Valero (Coords.), *La era de las TT.II.CC. en la nueva docencia* (pp. 109–188). Madrid: McGraw-Hill.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, *5*(1), 155–184.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main: Peter Lang.

Corpas Pastor, G., & Seghiri Domínguez, M. (2016). *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*. Frankfurt am Main: Peter Lang.

Costa, H., Corpas Pastor, G., & Seghiri, M. (2014). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. En *Translating and the Computer 36* (pp. 51–55).

Costa, H., Corpas Pastor, G., Seghiri, M., & Mitkov, R. (2015). iCorpora: Compiling, Managing and Exploring Multilingual Data. En G. Corpas Pastor, M. Seghiri Domínguez, R. Gutiérrez Florido & M. Urbano Mendaña (Eds.), *Nuevos horizontes en los Estudios de Traducción e Interpretación/ New Horizons in Translation and Interpreting Studies* (pp. 74–76). Ginebra: Tradulex.

Costa, H., Durán Muñoz, I., Corpas Pastor, G., & Mitkov, R. (2016). Compilaçao de Corpos Comparáveis Especializados: Devemos sempre confiar nas Ferramentas de Compilaçao Semi-automáticas? *Linguamática*, *8*(1), 3–19.

De Groc, C. (2011). Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. En *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (pp. 497–498).

Désilets, A., Farley, B., Stojanovic, M., & Patenaude, G. (2008). WeBiText: Building Large Heterogeneous Translation Memories from Parallel Web Content. *En Proceedings of Translating and the Computer*, 30 (pp. 27–28).

Esplà Gomis, M., & Forcada, M. (2010). Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. *The Prague Bulletin of Mathematical Linguistics*, *93*(1): 77–86.

Fletcher, W. H. (2004). Facilitating the Compilation and Dissemination of Ad Hoc Web Corpora. En G. Aston, S. Bernardini & D. Stewart (Eds.), *Corpora and Language Learners* (pp. 275–302). Ámsterdam: John Benjamins Publishing.

Fletcher, W.H. (2007). Concordancing the Web: Promise and Problems, Tools and Techniques. En M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 25–45). Ámsterdam: Rodopi.

Gatto, M. (2013). *Web as Corpus: Theory and Practice*. Londres/Nueva York: Bloomsbury.

Gutiérrez Florido, R., Corpas Pastor, G., & Seghiri, M. (2013). Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. En *10th Int. Conf. on Terminology and Artificial Intelligence (TIA'13)*, Paris, France.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7–36.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. En G. Williams & S. Vessier (Eds.). *Proceedings of the 11th EURALEX International Congress* (pp. 105–115).

López Rodríguez, C. I. (2016). Using Corpora in Scientific and Technical Translation Training: Resources to Identify Conventionality and Promote Creativity. *Cadernos de Traduçao*, *36*(1), 88–120.

López Rodríguez, C. I. & Buendía Castro, M. (2011). En busca de corpus online a la carta en el aula de traducción científica y técnica. *Trans-kom*, *4*(1), 1–22.

McEnery, T., & Wilson, A. (2011). *Corpus Linguistics*. Edimburgo: Edinburgh University Press.

Papavassiliou, V., Prokopidis, P., & Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. En *6th Workshop on Building and Using Comparable Corpora* (pp. 43–51).

Rossi, C., Frérot, C., & Falaise, A. (2016). Integrating Controlled Corpus Data in the Classroom: A case study of English NPs for French Students in Specialised Translation. En F. Alonso Almedia, L. Cruz García &. V. González Ruiz (Eds.), *Corpus-based Studies on Language Varieties* (pp. 167–190). Frankfurt am Main: Peter Lang.

Rychlý, P. (2007). Manatee/bonito-a modular corpus manager. En *1st Workshop on Recent Advances in Slavonic Natural Language Processing* (pp. 65–70).

Sánchez Gijón, P. (2009). Developing documentation skills to build do-it-yourself corpora in the specialized translation course. En A. Beeby, P. Rodríguez Inés & P. Sánchez-Gijón (Eds.), *Corpus use and translating. Corpus use for learning to translate and learning corpus use to translate* (pp. 109–127). Ámsterdam: John Benjamins Publishing.

Sánchez Ramos, M. M. (2017a). Metodología de corpus y formación en la traducción especializada (inglés-español): Una propuesta para la mejora de la adquisición de vocabulario especializado. *Revista de Lingüística y Lenguas Aplicadas*, *12*, 137–150.

Sánchez Ramos, M. M. (2017b). Compilación y análisis de un corpus ad hoc como herramienta de documentación electrónica en Traducción e Interpretación en los Servicios Públicos. *Estudios de Traducción*, *7*, 177–190.

Sanz Vicente, M. L. (2008). Propuesta metodológica basada en el uso de corpus bilingües para la elaboración de un diccionario de teledetección inglés-español. En D. Azorín Fernández (Ed.), *El diccionario como puente entre las lenguas y las culturas del mundo* (pp. 264–270). Alicante: Alicante: Universidad de Alicante y Fundación Biblioteca Virtual Miguel de Cervantes.

Seghiri, M. (2011). Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad. *Revista de Lingüística Teórica y Aplicada*, *49*(2), 13–30.

Seghiri, M. (2015). Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicos. En M. T. Sánchez Nieto (Ed.), *Corpus-based Translation and Interpreting Studies: From description to application* (pp. 125–146). Berlín: Frank & Timme.

Seghiri, M. (2017). Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores. *Babel*, *63*(1): 43–64.

Toledo Báez, M. C., & Martínez Lorente, R. (2018). Colocaciones, locuciones y compuestos sintagmáticos bilingües (español-francés) sobre diabetes en el corpus comparable *Cordiabicom*. *Panace@*, 47, 106–114.

Zanettin, F.; Bernardini, S., & Stewart, D. (Eds.). (2003). *Corpora in Translator Education*. Manchester: St. Jerome.